Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps:

An In-depth Look into National and State Reading and Math Outcome Trends

By

Jaekyung Lee
Graduate School of Education
State University of New York at Buffalo

Foreword by Gary Orfield

June 2006

This publication should be cited as:
Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends.* Cambridge, MA: The Civil Rights Project at Harvard University.

Additional copies of this report may be obtained from our website at:
http://www.civilrightsproject.harvard.edu

Produced with generous support from the Bill and Melinda Gates Foundation and the Charles Stewart Mott Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The author is very grateful to numerous individuals for their assistance with this report. Special thanks go to Gail Sunderman and Gary Orfield who contributed greatly to reviewing and editing this report. I am also very grateful to external reviewers, Doug Harris, Gene Glass, and Robert Linn who provided invaluable comments and suggestions on an earlier draft. I would also like to thank members of the Civil Rights Project's administrative team, including Jennifer Blatz and Lori Kelley who helped with the production of the report. Finally, I would like to acknowledge the contributions of research assistants, Jie Wang and Jeff Fox, for their help with data collection and analysis.

FOREWORD

The No Child Left Behind Act has hundreds of pages of complex provisions but simple and unambiguous goals. It embodies President Bush's promise to end the "soft racism of low expectations" by closing racial achievement gaps and bringing all students to proficiency within the next eight years. It creates unprecedented measurement of academic progress in two subjects (with science being added later) through mandated yearly tests in elementary and middle school and requires that all children from all racial and ethnic groups attain 100% proficiency. Schools are required, under threat of strict sanctions, to raise achievement each year in math and reading and to eliminate the achievement gap by race, ethnicity, language, and special education status.

The bipartisan bargain that led to the enactment of the law was designed around hope of dramatic educational progress spurred by large increases in federal aid and strict accountability. Many of the high poverty schools the law aimed to change had limited resources, poorly trained teachers, and instability of both student enrollment and staffing, making it very difficult to accomplish large educational breakthroughs without large increases in funds and major reforms. Unfortunately, after the first year, the promised resources were not provided but the very demanding standards remained in place. As it stands, the act can best be understood to represent the theory that large gains in achievement and equity can be quickly coerced out of the existing public school system without additional resources or long-term systemic reforms that take years to accomplish.

Given the bitter controversy over the wisdom and fairness of the basic structure of the law, which mandates reaching these 100% goals and only sanctions when they are not met, it is extremely important to determine whether or not the law is working on its own terms. With four years having passed since the law was first enacted, we must now ask whether the policies that have already labeled more than a fourth of all American schools as failures and initiated sanctions against them have succeeded.

This report concludes that neither a significant rise in achievement, nor closure of the racial achievement gap is being achieved. Small early gains in math have reverted to the preexisting pattern. If that is true, all the pressure and sanctions have, so far, been in vain or even counterproductive. The federal government is providing $412 million a year to help pay for part of the additional testing required by the law and many states claim that they are being forced to divert state funds to testing and other provisions they believe are unnecessary.

The reported state successes are artifacts of state testing policies which lead to apparent gains on state tests that simply do not show up on an independent national test, the National Assessment of Educational Progress (NAEP), the federal assessment system run by the Educational Testing Service, known as the "nation's report card." The study shows that virtually all states are reporting gains in achievement and many are reporting that the achievement gap is closing under their state assessment and proficiency systems, but that those gains are not related to gains independently measured by NAEP and are

largest in those states with the least demanding standards and the lowest thresholds for achievement.

On the issue of closing the gap for minority and poor children, a central goal of NCLB, there are also no significant changes since NCLB was enacted. Given the fact that we have 50 different state education systems, each with its own testing system, there is, of course, variation among the states. In terms of advocates of high stakes standards based reforms, however, it appears that gains were not related to the early adoption of those systems of accountability.

The Bush Administration and some of the policy's most fervent supporters, avoid this uncomfortable truth. Instead they have claimed that No Child Left Behind has produced a major breakthrough both in terms of achievement and in terms of closing the gap. The White House hails a 52% increase in spending on the key provision since 2001 and cites the research of the Education Trust, claiming that achievement is rising in 23 of 24 states studied and that in most of them the racial achievement gap had narrowed.

As the leader of a research project concerned about issues of racial equity, I believe that if there were evidence that these things were actually being accomplished it would be very important whatever one thought about some of the means being used to attain them. Unfortunately, these claims rest on misleading interpretations of flawed data as demonstrated in this new report.

The basic problem in the claims about gap closing is that state proficiency levels are simply a threshold measure and as more minority students cross that level, states claim that the gap is closing and achievement is rising. This is something like an athletic test that required people to jump over a two-foot barrier, but did not measure how much above the barrier anyone was jumping. Most students who were healthy would do this easily at the beginning. As others got some practice and training, more and more would meet this easy standard, but that would say nothing about whether the gap in jumping had closed since those who succeeded at first might well be jumping twice as high as they practiced, and the real gap would be getting wider, not narrower. It turns out that this gap widening is happening in many states, but not reflected in many state-testing reports.

This report indicates that the basic trends in both achievement gains are almost exactly what they were before the act became law—modest gains in math, flat achievement in reading. There are now modest gains on the NAEP in math, but the growth pattern is the same as that which existed before NCLB. Achievement on reading tests is basically unchanged. It shows that continuing the current trends will leave the nation very far from reaching the 100% proficiency goal. In Shakespearean terms, we've been experiencing a massive process "full of sound and fury, signifying nothing."

It is important to keep in mind that the NAEP does show substantial declines in racial achievement gaps in the l970s and early 1980s, when more of the civil rights and anti-poverty efforts of earlier reforms were still in operation The strict standards-based

reform effort that swept the county after the 1983 *A Nation at Risk* report has not shown similar benefits on achievement gaps.

Since the policy is little more than a theory about how to force change without any grounding in specific educational approaches or targeted resources to ensure that effective programs and supports are put into place (except in the special early reading programs), then if it does not succeed in improving scores on the NAEP, it certainly cannot be justified. This is after teachers and school leaders across the country have been put through tremendous stress, and a vast amount of money has been expended in developing and implementing tests states did not believe were necessary or well advised. Much more important, because of the high stakes attached to the tests, many millions of hours of class time have been committed to preparation for these tests. Under the law nothing about a school has counted in determining its success or failure except these math and reading tests. Obviously, under those circumstances, this pressure tends to drive other things out of the school day. If the policy fails to produce real gains even on those limited outcomes, it needs to be redesigned if the laudable goals are to be attained. Much could be learned from earlier Congressional efforts to tie Title I funds to multi-year, full school reform, to support the creation of magnet and other schools with less concentrated poverty, and to support school reform with broader anti-poverty efforts.

A combination of intense pressure for gains and a narrow focus on measurement means schools at risk of being branded as failures concentrate on moving those scores that will determine their fate. One way they do this is to focus more time on preparing for those particular tests at the expense of all the other outcomes that are not measured. For example, there is no accountability for whether or not students learn anything about American history and our democratic institutions. There is significant evidence that the students receive even less instruction than previously in subjects not tested and that excessive pressure can actually undermine another goal of the law—attracting highly qualified teachers to high poverty schools and holding them there. Our survey of teachers in California and Virginia school districts show most of those teachers believe that this narrowing has happened and a recent report from the Center for Education Policy shows that this pattern is widespread across the nation. An Associated Press study shows a sharp reduction of recess in a society with children whose physical fitness is declining. As schools are branded as failures, teachers in threatened schools narrow what they are teaching, eliminate instruction on subjects not on the tests, and tend to transfer out of the schools more rapidly. Things that help keep kids attached to the school experience like recess, arts and music, and career related training as well as extracurricular activities are reduced in pursuit of goals that are not being achieved.

Particularly in low income schools judged as failures, there often is a tendency to move into highly formulaic and rigidly programmed curriculum, boring to both students and teachers, and, worse yet, to spend time not on teaching their subjects but on drilling on test-taking strategies. Teachers have long tended to transfer out of low-income minority schools as they gain experience. Excessive test pressure tends to accelerate this process, compounding the schools' problems since experienced teachers are a precious resource for schools. We found this pattern in our teacher surveys.

One of the reasons why the overemphasis on test scores for accountability has such severe risks is that in schools which are threatened, there is an enormous incentive to invest in strategies such as teaching to the test, which creates the false impression of educational progress since what has been learned is more about the test and what is included on it, not deeper knowledge of the subject. In fact, there may be less in-depth understanding of the subject and less preparation for the higher order skills that come later and require a broader kind of preparation. When the means displace the end, when single measurements created by testing companies working with bureaucratic committees are treated as more important than all the other ways teachers assess their students, then it should not be surprising that we really have nothing that shows up on another measure of learning, even in those subjects where intense test preparation takes place.

Normally, after a vast reform, critics would be demanding results and proposing to restructure the program if gains were not forthcoming. I think that it is appropriate and necessary to assess the assessments, to evaluate the cost-benefit ratio of the current requirements and to propose changes in the assessment regime, the definition of adequate progress, and the time and resource required for deep reforms as the law is reauthorized.

Assuming that the data from the NAEP are accurate and that the policy so far has failed on its own terms and that there is no evidence that its goals will be attained if existing trends continue, what should a reader conclude? First of all, do not believe claims that are made from what is essentially meaningless data. Unless short-term changes in NAEP scores are compared to the trends that existed for years before NCLB became law as this study does, it is difficult to know if the changes are real improvements. Second, consider the possibility that the policy is simply wrong in its theory of educational change and needs to be modified. The best research suggests that school reform takes time, that investments must be made in curriculum and instruction, and that sustaining educational improvement in high poverty schools is difficult at best. It is much easier to attract and hold teachers in schools where they are needed by rewards when they make a difference than with constant threats. Third, we must, of course, concede that it is possible that the longer-term results of NCLB will be less disappointing than those in the first several years of the project. One would hope that something in which this much treasure and effort has been invested would show some significant results eventually. Even then, of course, one would have to evaluate those possible gains against the costs of the policy.

As No Child Left Behind will soon be up for reauthorization, I believe that this research shows, at the minimum, that the theory of test-driven change underlying NCLB is too simple, that the goals must be more realistic, and that the thought of accomplishing much without the promised resources is probably unrealistic. Congress should be open to other ideas and should listen to educators who have actually accomplished major breakthroughs and to researchers outside the Washington advocacy networks who have actually documented what kinds of reforms can work, how much they can accomplish, what is a reasonable theory of change, and what kind of time and resources are needed to realize its potential. The goals of raising achievement and lowering gaps are very good ones, and the data provided by NCLB is essential, but policy makers must be ready to

critically examine why so little has been accomplished, why officials are making misleading and inaccurate claims, and what can be done to use the invaluable data and focus created by the Act to begin to actually accelerate progress toward those objectives. If we take these results as showing the need for substantial mid-course corrections, not as an attack on the goals of good teachers in poor schools and meaningful accountability for real progress by all groups of students, I believe that we could begin to actually close the gaps again, as we were doing a generation ago. I believe that educators across the country would be eager to work with Congress and the Administration in making the needed changes so we could produce real gains that would show up on whatever test the students took.

Gary Orfield

EXECUTIVE SUMMARY

Achievement gaps constitute important barometers in educational and social progress. The National Assessment of Educational Progress (NAEP), the so-called nation's report card of student achievement, provides information on the achievement gaps among different racial and socioeconomic groups in core academic subjects. In the 1990s, there were significant setbacks in the national progress toward narrowing the achievement gaps. Few states were able to improve the average achievement and narrow the gaps simultaneously. The No Child Left Behind Act of 2001 (NCLB), aims at ensuring both academic excellence and equity by providing new opportunities and challenges for states to advance the goal of closing the achievement gap.

Research findings on the effects of high-stakes testing on improving academic performance have been mixed, generating controversy over the policy's usefulness. While NCLB builds upon the alleged success of first generation states, which adopted test-based accountability systems prior to NCLB, assessing its impact requires more rigorous scrutiny of new evidence from NAEP and state assessment results. Although NCLB establishes state assessments as the basis for NCLB accountability, NAEP can play a confirmatory role as an independent assessment to validate the state test results. Previous studies are limited because they rely on states' own assessment results only "after" NCLB was adopted to show changes in achievement. Any change we see in student achievement after NCLB may reflect a continuing trend that occurred before NCLB. It remains to be examined whether and how recent NAEP reading and math assessment trends in average achievement as well as racial and socioeconomic achievement gaps are systematically related to federal and state accountability policies under NCLB.

This study offers systematic trend analyses of NAEP national and state-level public school fourth and eighth graders' reading and math achievement results during pre-NCLB (1990-2001) and post-NCLB (2002-2005) periods. It compares post-NCLB trends in reading and math achievement with pre-NCLB trends among different racial and socioeconomic groups of fourth and eighth graders from across the nation and states. National and state progress toward closing racial and socioeconomic achievement gaps are evaluated not only in terms of their success in reducing the test score gaps but also in terms of reducing each subgroup's chance of failing to meet desired performance standards. Further, it provides new evidence on the impact of state accountability policy on the achievement gap trends and the discrepancies between NAEP and state assessment results.

Key Findings

- NCLB did not have a significant impact on improving reading and math achievement across the nation and states. Based on the NAEP results, the national average achievement remains flat in reading and grows at the same pace in math after NCLB than before. In grade 4 math, there was a temporary improvement right after NCLB, but it was followed by a return to the pre-reform growth rate.

Consequently, continuation of the current trend will leave the nation far behind the NCLB target of 100 percent proficiency by 2014. Only 24 to 34 percent of students will meet the NAEP proficiency target in reading and 29 to 64 percent meeting that math proficiency target by 2014.

- NCLB has not helped the nation and states significantly narrow the achievement gap. The racial and socioeconomic achievement gap in the NAEP reading and math achievement persists after NCLB. Despite some improvement in reducing the gap in math right after NCLB, the progress was not sustained. If the current trend continues, the proficiency gap between advantaged White and disadvantaged minority students will hardly close by 2014. The study predicts that by 2014, less than 25 percent of Poor and Black students will achieve NAEP proficiency in reading, and less than 50 percent will achieve proficiency in math.

- NCLB's attempt to scale up the alleged success of states that adopted test-driven accountability policy prior to NCLB, so-called first generation accountability states (e.g., Florida, North Carolina, Texas) did not work. It neither enhanced the first generation states' earlier academic improvement nor transferred the effects of a test-driven accountability system to states that adopted test-based accountability under NCLB, the second generation accountability states. Moreover, both first and second generation states failed to narrow NAEP reading and math achievement gaps after NCLB.

- NCLB's reliance on state assessment as the basis of school accountability is misleading since state-administered tests tend to significantly inflate proficiency levels and proficiency gains as well as deflate racial and social achievement gaps in the states. The higher the stakes of state assessments, the greater the discrepancies between NAEP and state assessment results. These discrepancies were particularly large for Poor, Black and Hispanic students.

While one should interpret the findings from this study about the impact of NCLB cautiously, the study has implications for the debate about reauthorization. NCLB requires adequate yearly progress of all groups of students toward the state proficiency target. The report demonstrates how, over the past few years since NCLB's inception, state assessment results show improvements in math and reading, but students are not showing similar gains on the NAEP—the only independent national test. If we continue the current policy course, academic proficiency is unlikely to improve significantly, but it is possible that the state assessment will continue to give a false impression of progress, shortchanging our children and encouraging more investment into a failed test-driven accountability reform policy. This problem can be more serious for schools that serve predominantly disadvantaged minority students. NCLB has shortchanged those schools with under-funded mandates and an over reliance on sanctions rather than a focus on capacity building. While failure is not an option in education, it is important to acknowledge the limitations of the current policy and find solutions to problems that may have impeded national and state progress towards academic excellence and equity.

PART 1: INTRODUCTION

Achievement Gap Trends

Achievement gaps constitute important barometers in educational and social progress. The National Assessment of Educational Progress (NAEP), the so-called nation's report card of student achievement, provides information on the achievement gaps among different racial and socioeconomic groups in core academic subjects.[1] Racial and socioeconomic achievement gaps narrowed substantially in the 1970s and early 1980s. During the 1970s, education and social policies worked to narrow the achievement gap by guaranteeing a minimally adequate level of achievement for minorities through compensatory education, minimum competency testing, school desegregation, equalization of school funding, the war on poverty, and affirmative action. As the focus of education policy has shifted from equity to excellence during the last two decades, there is a potential tension between academic excellence and equity (Bracey, 2002; O'Day & Smith, 1993). In the 1990s, racial achievement gaps stopped narrowing or began to widen, signaling setbacks in the progress the nation made toward educational equity (Lee, 2002).[2]

States were not effective in addressing educational inequalities and achievement gaps in the 1990s (Braun, Wang, Jenkins, Weinbaum, 2006; Lee & Wong, 2004). A report to the National Education Goals Panel notes that states made little progress in narrowing the persistent gap in mathematics achievement between White and minority students and between Poor and better-off students during the 1990s, despite overall gains in achievement scores on the NAEP (Barton, 2002). The gaps remain substantial as of 2005. For example, the 2005 NAEP report not only shows that the percentage of Black and Hispanic students performing at or above the Proficient level in mathematics is much lower than that of their White peers (47 % for Whites vs. 13 % for Blacks and 19 % for Hispanics at grade 4; 39 % for Whites vs. 9 % for Blacks and 13 % for Hispanics at grade 8), but it also shows that a large majority of Black students fail to meet the proficiency standard. Simply reducing disparities in test scores is not sufficient without also improving the percentage of low-achieving students and disadvantaged minority groups that perform at or above the NAEP proficiency level.

The No Child Left Behind Act of 2001 (NCLB) aims at ensuring both academic excellence and equity by providing new opportunities and challenges for states to advance the goal of closing the achievement gap. It relies on high-stakes testing to ensure that schools make adequate yearly progress (AYP) toward the goal of 100% proficiency by 2014. Researchers and educators have raised concerns about the negative consequences of NCLB's test-based accountability

---

[1] For example, there is a documented achievement gap in mathematics between White students and minority students in the U.S., particularly socio-economically disadvantaged Black and Hispanic students; an average Black high school graduate's standardized math test score can be as low as that of an average White 8th grader.

[2] For diverse perspectives on the issue of closing the achievement gap, see Jencks & Phillips (1998), Peterson (2006), and Rothstein (2004).

and its uniform AYP requirement, including its potential to perpetuate or exacerbate existing racial, economic, or geographic inequalities among schools (Kim & Sunderman, 2005; Lee, 2003; Lee, 2004; Linn, 2003; Sunderman, Kim, & Orfield, 2005). Also, education advocates, state education officials, and some members of Congress were concerned about unfunded NCLB mandates and called for more serious federal efforts to accomplish the original intent of the law (Mathis, 2003; NAACP, 2005; NSBA, 2006).[3]

Does external, test-driven accountability policy enhance or hinder academic excellence and equity? The research findings on the effects of high-stakes testing on improving academic performance have been mixed, generating controversy over the policy's usefulness (Amrein & Berliner, 2002; Carnoy & Loeb, 2002; Grissmer & Flanagan, 1998; Hanushek and Raymond, 2004; Lee, in press-b; Lee & Wong, 2004; Nichols, Glass, & Berliner, 2006; Raymond & Hanushek, 2003; West & Peterson, 2005).[4] The case that drew the most attention was Texas, where the evidence on the impact of high-stakes testing was highly contested (Carnoy, Loeb, & Smith, 2001; Grissmer & Flanagan, 1998; Grissmer, Flanagan, Kawata, & Williamson, 2000; Haney, 2000; Skrla, Scheurich, Johnson, & Koschoreck, 2004; Valencia, Vanezuela, Sloan, & Foley, 2004). While NCLB builds upon the alleged success claimed by some of these earlier studies of states that had adopted accountability policies prior to NCLB (first-generation accountability states such as Florida, North Carolina and Texas), assessing its impact requires a more rigorous scrutiny of new evidence from NAEP and state assessment results from across the nation and states. Left unexamined is whether and how the recent NAEP reading and math assessment trends in average achievement as well as racial and socioeconomic achievement gaps are systematically related to new federal and state accountability policies under NCLB.

Mixed Reactions to NAEP Reports on Post-NCLB Reading and Math Achievement Trends

NAEP can provide timely information to states regarding their students' achievement against high performance standards in core subject areas. In light of the controversies about the impact of NCLB on student outcomes, many people were anxious to see the NAEP 2005 results, which reported the national and state average reading and math performance from testing nationally representative samples of more than 300,000 4th and 8th graders (Perie, Grigg, & Donahue, 2005 for reading; Perie, Grigg, & Dion, 2005 for math). Since the U.S. Department of Education released the NAEP 2005 reports

[3] In 2005, the State of Connecticut sued the U.S. Department of Education over insufficient funding and support from the federal government to help the state meet the testing provisions of NCLB (Connecticut v. Spellings).

[4] While many studies examined the average policy effect for all students, only a few (e.g., Carnoy & Loeb, 2002; Hanushek & Raymond, 2004; Lee & Wong, 2004) disaggregated the results by racial subgroups and explored potential accountability policy effects on racial achievement gaps. For recent comprehensive review of the literature on accountability policy impact, see Harris & Herrington (2006) and Lee (2006).

on October 19, 2005, reactions to these reports varied. The U.S. Department of Education newsletter, *The Achiever* (Vol. 4 No. 12, Nov/Dec 2005), noted, "overall math scores for both groups (4th and 8th graders) rose to all-time highs, and fourth-grade reading scores matched the all-time record." The U.S. Secretary of Education attributed credit for such gains to NCLB, who claimed "These results, like the long-term July data, confirm that we are on the right track with No Child Left Behind, particularly with younger students who have benefited from the core principles of annual assessment and disaggregation of data." (*The Achiever*, p. 2). Critics of standardized testing interpreted the 2005 NAEP results more negatively. The National Center for Fair and Open Testing (FairTest, 2005, October 19) commented on the NAEP 2005 report in its press release: "Flatline NAEP scores show the failure of test-driven school reform. No Child Left Behind has not improved academic performance." FairTest claimed that "NAEP reading scores were essentially unchanged from 2002 to 2005 at grade 4 and declined markedly at grade 8." FairTest also pointed out that "math scores did not increase at a significantly faster rate than in the 1990s, well before most high-stakes exams for elementary and middle school were put in place."

The different interpretations of the same results may be attributed partly to differences in the time frame used to analyze changes in the test results and different ways of evaluating the policy impact. To understand whether short-term improvements in NAEP scores can be attributed to NCLB, we need to assess any short-term changes in scores within a longer-term time frame. This will allow us to determine whether NCLB had a significant effect on academic growth or if the changes were the continuation of a growth pattern that began before NCLB and continued after its passage. In addition, changes in NAEP scores need to be analyzed within the broader context of testing and accountability policy.

Discrepancies between NAEP and States' Assessment Reports on Reading and Math Achievement Trends

State assessments are the basis for states' educational accountability decision-making under NCLB. Although NCLB does not prescribe a role for NAEP in making state accountability decisions, it does specify using NAEP scores to confirm state test results, to evaluate the rigor of state standards, and to show whether states are making progress in improving student achievement and reducing the achievement gap among concerned subgroups of students (Ad Hoc Committee on Confirming Test Results, 2002; Henderson-Montero, Julian, & Yen, 2003). Previous comparisons of NAEP and state assessment results showed significant discrepancies in the level of student achievement, as well as in the size of statewide achievement gains (Klein, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Lee, in press-a; Linn, Baker, & Betebenner, 2002). The percentages of students reaching the Proficient level tend to be generally lower on NAEP than on state assessments. These results suggest that, for many states, NAEP proficiency levels are more challenging than the states' own (National Education Goals Panel, 1996). Since state standards vary widely in relationship to NAEP standards, it raises questions about the generalizability of gains reported on a state's own assessment, and hence about the validity of claims regarding student achievement (Linn, 2000).

While several studies have attempted to examine the impact of NCLB on student achievement, they are limited because they use a single measure of achievement only "after" NCLB was adopted. Any change we see after NCLB may reflect a continuing trend that occurred before NCLB. Any changes that were clearly on track before NCLB should not be credited to the new law. While it is important to maintain the pace of improvement, it is inappropriate to credit NCLB for improving achievement if the law did not accelerate the pace. States tend to show progress on their own standards regardless of whether or not it transfers into progress independently measured by NAEP. For example, a report by the Education Trust (2004) on post-NCLB achievement trends relied solely on states' own assessment results. The report examined short-term changes in average achievement in state reading and math assessment results and changes in racial and economic achievement gaps after NCLB (from 2002 to 2004). The findings of this report suggest that the improvements in performance were positive but that narrowing the gap was slow. A follow-up report by the Education Trust (2006) takes a more comprehensive look into post-NCLB changes across grade levels (from 2003 to 2005) and finds more positive results at the elementary education level than at the secondary level. According to a report by the Center on Education Policy (2006), national survey results show that scores on state tests have risen in a large majority of states and school districts. That report credited school district policies and programs as more important contributors to these gains than the NCLB AYP requirements. Despite these earlier findings, real full-scale impact of NCLB on student achievement remains to be examined.

Design and Organization of Studies in the Report

This study offers a systematic analysis of trends on national and state-level public school students' reading and math achievement using data from NAEP (see Appendix A for descriptions of data and methods). Particular attention is paid to the achievement gap among racial and socioeconomic groups of students. Racial achievement gaps focus on the gaps between Blacks and Whites and between Hispanics and Whites. Socioeconomic achievement gap is measured by the gap between Poor (those eligible for free or reduced-priced school lunch) and Nonpoor student groups. National and state progress toward closing achievement gaps are evaluated not only in terms of their success in reducing the achievement gap in test scores but also in terms of reducing each subgroup's chance of failing to meet desired performance standards.

PART 2 reports findings from trend analyses that explore the effects of NCLB accountability policy on student achievement outcomes. Trend analyses involves fitting statistical models with estimates of pre-NCLB and post-NCLB change parameters based on a series of measurements on key outcome criteria obtained at periodic intervals before and after NCLB. It enables the evaluator to interpret the pre-to-post-NCLB changes by showing whether the achievement gains after NCLB are a continuation of earlier trends or whether they mark a decisive change. It is important to look at both average scores over time and trends in the achievement gap, since narrowing the gap without improving

average scores is not progress. In PART 2, this study compares the Pre-NCLB Period (1990 – 2001) with the Post-NCLB Period (2002 – 2005).[5]

Figure 1 illustrates three potential growth patterns that may result from NCLB policies. When NCLB has a significant positive effect, the performance trajectory will shift upwards with a marked increase in the growth rate (Scenario A in Figure 1). In this case, we expect sustained positive gain after NCLB so that post-NCLB growth rate is significantly greater than pre-NCLB growth rate. When NCLB has a significant negative effect, the performance trajectory will shift downwards with a marked decrease in the growth rate (Scenario C in Figure 1). When NCLB has no effect at all, a preexisting growth pattern will continue (Scenario B in Figure 1). In this case, we expect no change in the slope so that pre-NCLB and post-NCLB growth rates remain the same.

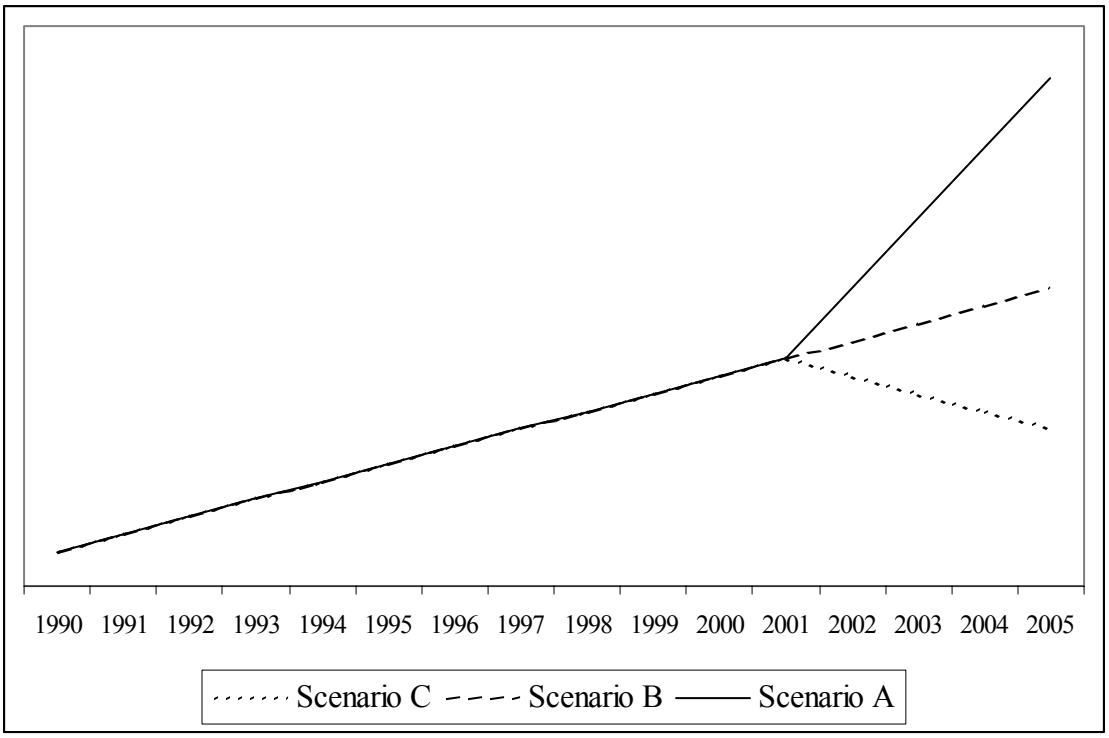

Figure 1: Hypothetical Achievement Trends Before and After NCLB: Positive Effect (Scenario A), No Effect (Scenario B), Negative Effect (Scenario C)

[5] Although the national achievement trends are simply divided into the two time periods, that is, pre-NCLB (1990-2001) vs. post-NCLB (2002-2005) for the sake of analysis, the pre-NCLB trend may reflect the influence of a precursor to NCLB, the Improving America's School Act (IASA) of 1994. It needs to be noted that IASA required states to develop assessment systems for measuring AYP, but NCLB substantially strengthened the scope and intensity of test-driven external accountability provisions by targeting all schools as opposed to Title 1 schools only as well as imposing more stringent requirements (e.g., meeting AYP for all subgroups) with real threats of punitive and corrective actions.

If the analysis were to find a distinctive effect of NCLB, it could not, of course, be attributed to just one part of NCLB such as high-stakes testing and school accountability. Other policy initiatives under NCLB, such as teacher qualification and parental involvement policies, and an initial influx of new federal funds, may have influenced the trends as well.[6] It is not possible to sort out the effect of one particular policy component from such an omnibus legislation. Moreover, some states that had high-stakes testing accountability prior to NCLB continued their own policies along with NCLB, thus creating a dual accountability system. By the same token, the pre-NCLB period is not free of similar types of interventions since some of the states already had their own accountability systems in place.

In PART 3, the study addresses variation among states in NAEP growth rates by taking into account their accountability policy history prior to NCLB. States which did not have high-stakes accountability policies before NCLB and were only exposed to the influences of external accountability under NCLB are compared with states that were active in test-driven accountability policy prior to NCLB. This analysis compares differences in both pre-NCLB and post-NCLB growth rates between two groups of states (Appendix B). States that adopted accountability polices before NCLB are called "first-generation" accountability states and include Kentucky, Maryland, North Carolina, California, Florida, New York, and Texas (Mintrop & Trujillo, 2005). States that never initiated statewide accountability reform before NCLB are called "second-generation" accountability states for the sake of distinction, whether they embraced NCLB or not.

Lawmakers did not intend that NCLB would replace a state's preexisting accountability policy where a parallel system already existed but rather would function as an add-on to enhance or augment state policy.[7] States with strong accountability systems may be better prepared to embrace and implement NCLB reform policy since implementation theory predicts stronger implementation fidelity among people who are accustomed to the intervention. No matter what real impact NCLB may have had on first generation states, the primary target of NCLB may have been second-generation states—those states where test-driven external accountability was new and where NCLB attempted to extend accountability modeled after the alleged success stories of some first-generation states such as Texas and North Carolina. By this logic, states with no exposure to high-stakes testing prior to NCLB are more likely to experience the effect of this new intervention by accelerating the pre-NCLB growth rate.

---

[6] For the theory of action for educational accountability policy, see Adams and Kirst (1999), Elmore and Fuhrman (1995), Fuhrman (1999), Lee and Wong (2004), Newmann, King, & Rigdon (1997), O'Day (2002).

[7] The existence of dual accountability systems and interactions between federal and state policies under NCLB poses methodological challenges for the analysis of post-NCLB data. Hanushek and Raymond (2004) point out the problem in that the implementation of NCLB essentially precludes analysis of further impacts of overall accountability systems by eliminating comparison group of states without accountability systems but at the same time the possibility that the continuation of individual states' own locally developed schemes affords comparison of the impacts of alternative types of accountability systems. NCLB also provides funding to support school improvement programs, and the interaction between NCLB accountability policy and preexisting school reform strategies such as Comprehensive School Reform (CSR) may affect the policy impact (LeFloch, Taylor, & Thomsen, 2006).

Hierarchical linear modeling (HLM) method, growth curve modeling, was used to examine trends in achievement in first and second generation states, recognizing that comparing two nonequivalent groups poses a threat to the validity of causal inferences about NCLB. The initial performance status gap (i.e., test score difference in 1990) reflects the fact that lower-achieving states were more active in adopting test-driven external accountability policies prior to NCLB. Until NCLB, states adopting test-driven accountability systems (first generation states) were expected to make greater test score gains than states not adopting these types of reforms (second generation states). After NCLB, both were likely to make about the same rate of growth. Consequently, second generation states may make greater pre- to post-NCLB progress in test score gains than their first generation counterparts. Latent variable HLM analysis was used to control for the effect of initial status on pre-NCLB gain and also the effect of pre-NCLB growth rate on post-NCLB change.

Finally, the study examines discrepancies between NAEP and states' own assessment results, explores factors, such as the degree of high-stakes testing, that might account for variations among states in these patterns, and discusses the policy implications of these findings. Previous studies that compared state assessments with NAEP scores were often restricted to a single state and did not systematically examine patterns across multiple grades and subjects from all states. In particular, those prior studies did not often look into possible differences between NAEP and state assessments in their estimation of the achievement gaps, an important indicator of state performance in educational equity (for exceptions, Lee, in press-a; Linton & Kester, 2003).

In light of these concerns, we need to examine whether and how both NAEP scores and states' own student assessments can be used to inform us of statewide academic performance. We also need to examine whether national and state assessments produce consistent results over time, particularly before and after NCLB. This requires a systematic comparative analysis of national and state student assessment data, specifically data on the proficiency levels of students, the achievement gaps among different racial groups of students, and their academic progress. The objective of the analysis presented in PART 4 is to investigate discrepancies between national and state assessment results at the state level and to explain interstate variations in the discrepancies.

One should interpret the findings from this study cautiously. This evaluation of the impact of NCLB on improving student achievement and narrowing the achievement gaps uses currently available NAEP data. Analysis over a longer time period may produce different results. Since there are only a few years of NAEP or state assessment data available for post-NCLB analysis, it may be premature to evaluate the full impact of NCLB as the policy sets 2014 as the deadline for states to meet it performance targets. Secondly, this analysis of repeated cross-sectional data confounds the policy effect and the cohort effect.[8] To address the possible influence of the cohort effect, we would also

[8] Concern about the cohort effect arises from the possibility that changes in the demographic compositions of NAEP student samples coincide with the policy intervention and both policy and demographic forces influence achievement trends at the same time (see Appendix A).

need to analyze demographic changes (racial and economic composition of successive cohort groups), something that requires long-term data.

With these caveats in mind, the findings of this report still have implications for NCLB, and test-driven external accountability policy in particular, as we approach the debate about reauthorization. Findings from the following series of extensive statistical data analyses are expected to provide policymakers and practitioners with useful information on the national and state trends in achievement gaps and can be used to help them develop policies that improve both equity and excellence. This study is also expected to inform policymakers of the discrepancies between NAEP and states' own assessment results and the importance of using multiple measures for accountability.

PART 2: NATIONAL TRENDS IN NAEP

Using NAEP reading and math assessment data from 1990 to 2005, the following analysis examines national trends in 4th and 8th grade students' academic growth before and after NCLB. The first section uses scale scores, that is, scores that summarizes the overall performance attained by a group of students on the NAEP, to show how average reading and math scores have changed over time. The second section examines changes in the percentage of students reaching the NAEP proficiency level, that is, the percentage of students, either in the total population or in a subgroup, that meet or exceed the NAEP proficiency level.[9]

National NAEP Reading and Math Scale Score Trends

Trends in the Average Achievement: Trends in national average reading and math gains on the NAEP are shown in Figure 2 and 3 respectively.[10] When comparing the average gains in reading achievement scores before NCLB with gains made after NCLB, we find no differences in the amount of gains made in grade 4 reading scores (Figure 2). Reading scores did not improve after NCLB and made only modest improvements prior to NCLB. In grade 8, there was a marked decline in average reading scores after NCLB compared to the pre-NCLB period. In contrast, math achievement scores showed significant improvement both before and after NCLB in both grades (Figure 3). However, the post-NCLB achievement growth pattern was not different from the pre-NCLB growth patterns.

---

[9] There are three achievement levels on the NAEP: Basic, Proficient, and Advanced. The achievement levels were authorized by the NAEP legislation and adopted by the National Assessment Governing Board (NAGB). They are collective judgments, gathered from a broadly representative panel of teachers, education specialists, and members of the general public, about what students should know and be able to do relative to a body of content reflected in the NAEP assessment frameworks. For reporting purposes, the achievement level cut scores for each grade are placed on the traditional NAEP scale resulting in four ranges: below Basic, Basic, Proficient, and Advanced. For example, the Proficient level of 8th grade math achievement was set at a score of 299 on a 0 to 500 NAEP scale, and eighth-grade students performing at this level should exhibit evidence of conceptual and procedural understanding in math (Mullis et al., 1993).

[10] NAEP results with accommodation permitted are shown for 1998-2005 years in reading and for 1996-2005 years in math. All prior assessments were done without accommodation.

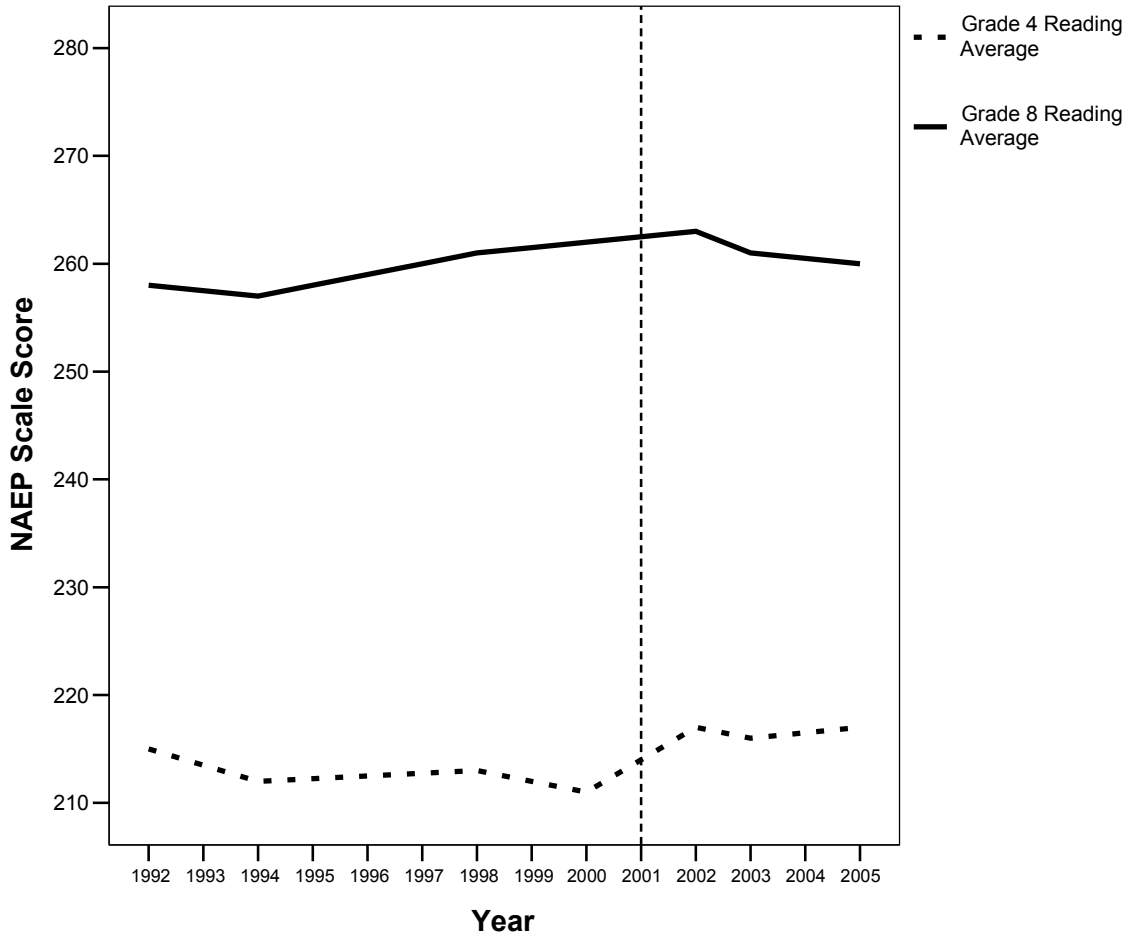

Figure 2: 1992-2005 NAEP Average Score Trends in Grade 4 and Grade 8 Reading

The average reading score gain from 2002 to 2005 (post-NCLB) was null for grade 4 and minus 3 points for grade 8. It is worth noting that although there was a temporary increase between 2000 and 2002 in the grade 4 average reading score, it was followed by a return to the pre-reform growth rate (Figure 2). The average amount of national public school "3-year" reading gain prior to NCLB, during the period of 1992-2002 (when NAEP data was available), was .6 point (2 points for 10 years multiplied by 0.3) at grade 4 and 1.5 points (5 points for 10 years multiplied by 0.3) at grade 8. Thus, the 3-point drop between 2002-05 in grade 8 may signify a marked decline in comparison with the pre-NCLB period.

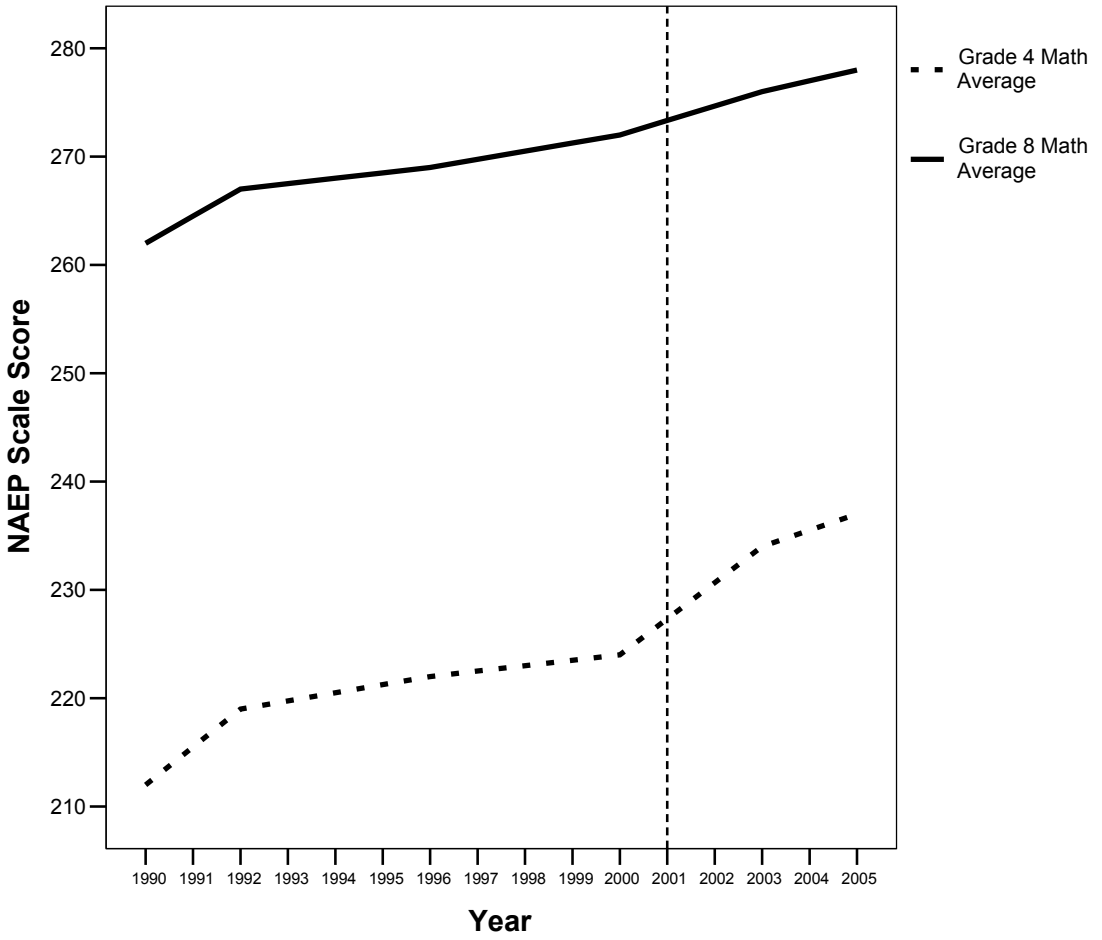

Figure 3: 1990-2005 NAEP Average Score Trends in Grade 4 and Grade 8 Math

The average math gain between 2003 and 2005 was 3 points for 4th grade and 2 points for 8th grade. Although there was a temporary increase between 2000 and 2003 in grade 4 average math score, it was followed by a return to the pre-reform growth rate (Figure 3). The average amount of national public school 2-year math gain prior to NCLB, during the period of 1990-2000 (when NAEP data was available), was 2.4 points (12 points for 10 years multiplied by 0.2) at grade 4 and 2 points (10 points for 10 years multiplied by 0.2) at grade 8. So the 3-point and 2-point gain between 2003-05 in grades 4 and 8 respectively are statistically significant per se but not different than the earlier achievement growth pattern.

Trends in the Achievement Gap: The racial achievement gap persists after NCLB. The achievement gap between White and Black students and between White and Hispanic students remained unchanged in both reading and math in both grades 4 and 8. The only significant change was a small reduction in the achievement gap between White

and Hispanic students in grade 8 math. Likewise, the gap between Poor and Nonpoor students remained.

Figure 4 and Figure 5 show the NAEP trends in the White-Black achievement gap between 1992 and 2005 in reading and between 1990 and 2005 for math. For instance, the average Black–White math score gap for eighth graders changed from 32.9 in 1990 to 33.4 in 2005, and that change was not statistically significant (Figure 5). Although there was a temporary drop between 2000 and 2003 in the grade 8 math White-Black gap, the gap leveled off afterwards. There was more progress in reducing the Black–White gap in math for fourth graders, which narrowed from 31.2 in 1990 to 26.0 in 2005. However, the pattern of post-NCLB change in the gap was not significantly different from its corresponding pre-NCLB trend.

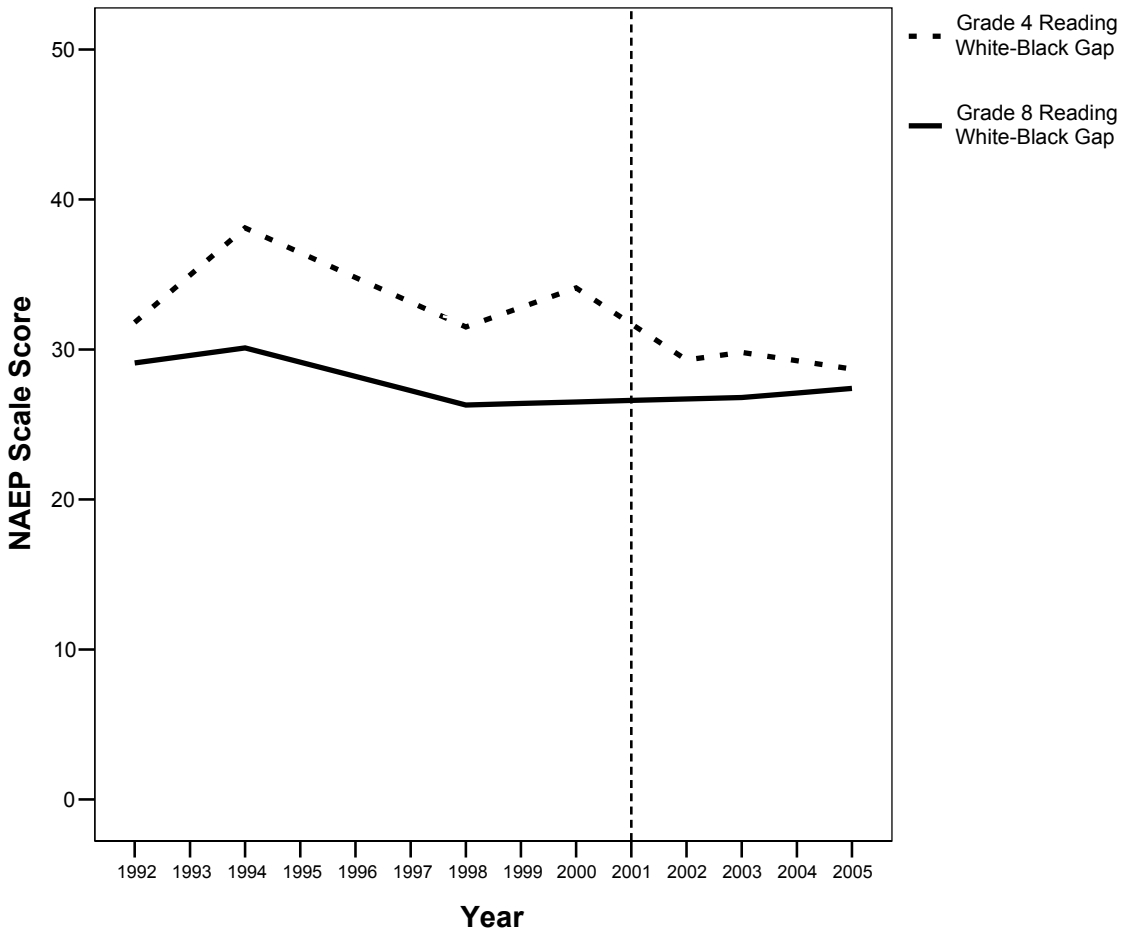

Figure 4: 1992-2005 NAEP White-Black Gap Trends in Grade 4 and Grade 8 Reading

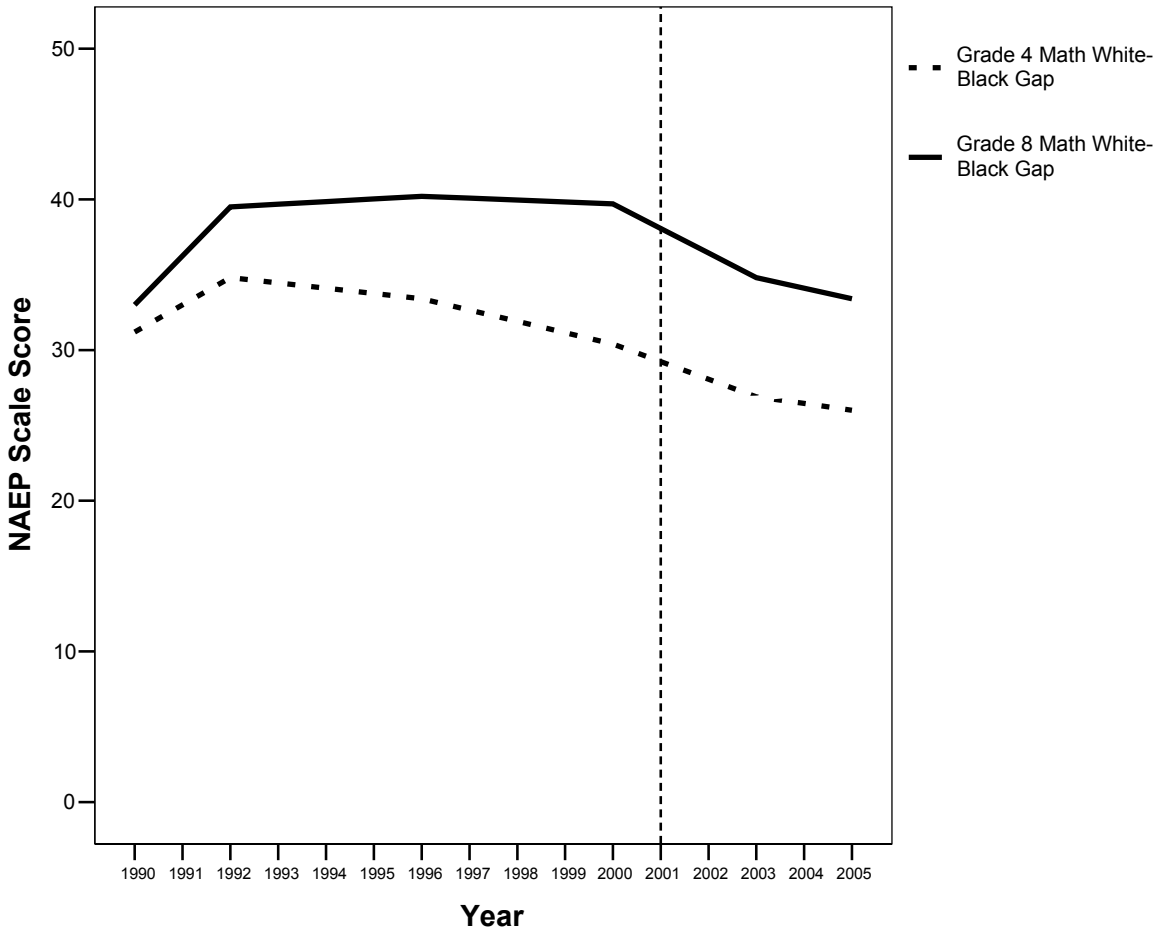

Figure 5: 1990-2005 NAEP White-Black Gap Trends in Grade 4 and Grade 8 Math

Similarly, Hispanic–White reading and math achievement gaps have hardly changed over the period. Figure 6 and Figure 7 show the NAEP trends in the White-Hispanic achievement gap between 1992 and 2005 in reading and between 1990 and 2005 for math. For example, the average Hispanic–White math score gap for eighth graders was 24.4 in 1990 and 26.5 in 2005 (Figure 7). The cumulative amount of the gap change during the 1990-2005 period was not significant, except that there was significant reduction between 2000 and 2005 in the grade 8 math White-Hispanic gap. The average Hispanic–White math score gap for fourth graders was 19.4 in 1990 and 20.6 in 2005 (Figure 7). Although there was a temporary drop between 2000 and 2003 in the grade 4 math White-Hispanic gap, the gap leveled off afterwards. A similar pattern is found in grade 4 reading. Consequently, the White-Hispanic gap in reading and math has returned back to its baseline level by 2005.

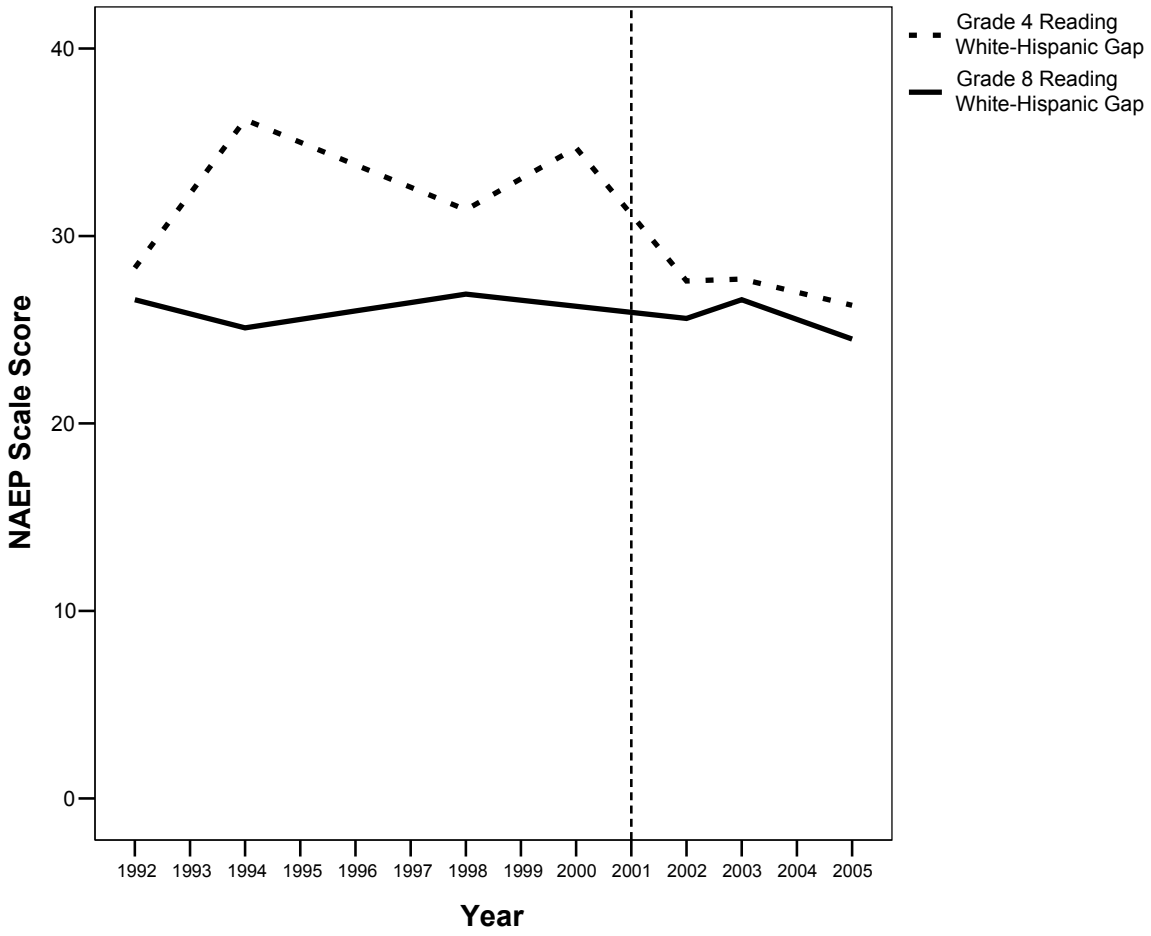

Figure 6: 1992-2005 NAEP White-Hispanic Gap Trends in Grade 4 and Grade 8 Reading

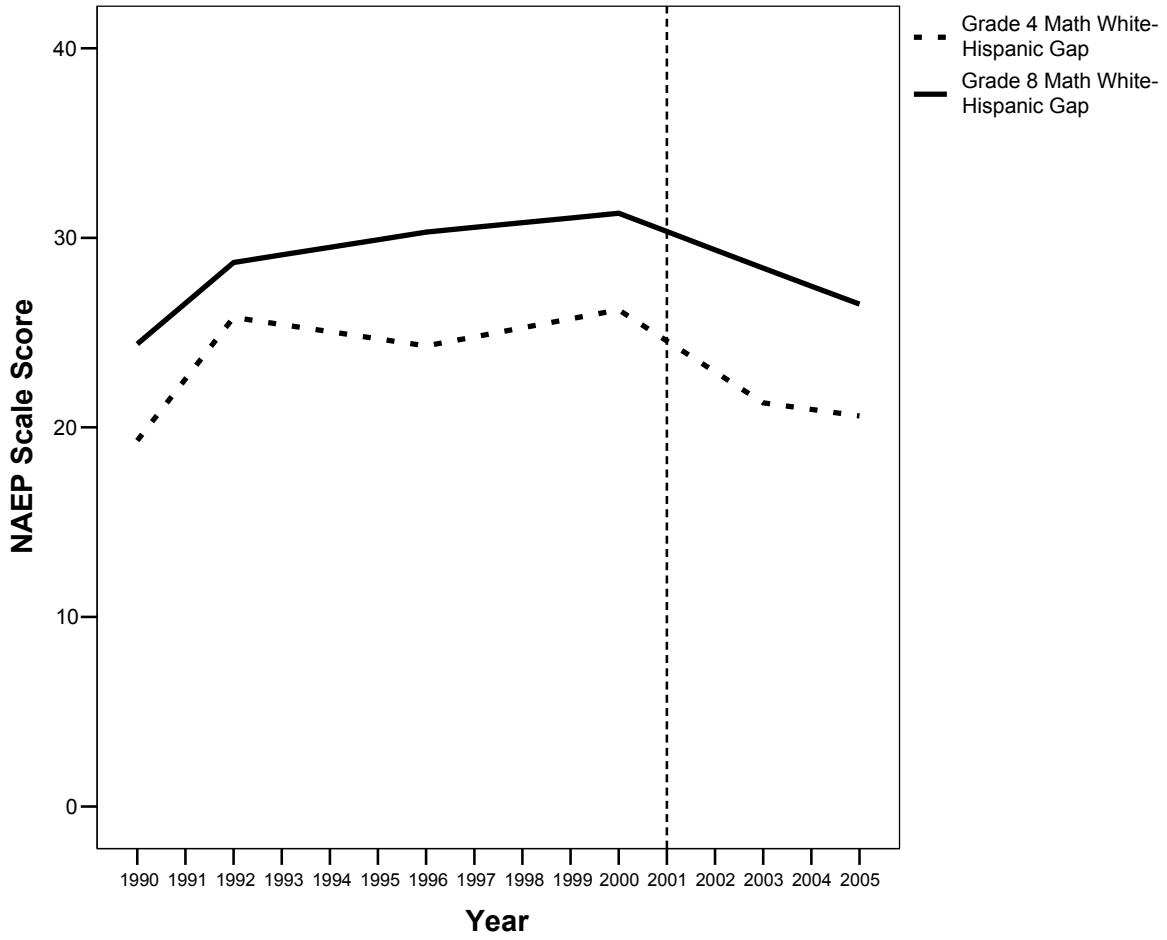

Figure 7: 1990-2005 NAEP White-Hispanic Gap Trends in Grade 4 and Grade 8 Math

By and large, the racial achievement gap in national public schools persists after NCLB. The White-Black and White-Hispanic gaps among 4th and 8th graders did not narrow significantly between 2002 and 2005 in reading and between 2003 and 2005 in math. The racial gap in reading remained about the same between 2002 and 2005 at both grade 4 and grade 8; the one-point change was not only statistically insignificant but also it is much smaller than the 5 point reduction of the gap made during the 2000-2002 period. The White-Black and White-Hispanic reading gaps at grade 4 increased in the early 1990s and then decreased in the late 1990s and by 2002 (prior to NCLB). The White-Black and White-Hispanic gaps remained unchanged at grade 8 throughout the 1992-2005 period. The racial gap change in math between 2003 and 2005 is also not significant. The only significant change, albeit small, is a two-point reduction in the achievement gap between White and Hispanic students in grade 8 math scores.

As shown in Figure 8 and Figure 9, the poverty gap also did not change significantly in both reading and math at grades 4 and 8. For example, the achievement

gap in NAEP eighth-grade math between Poor and Nonpoor students remained unchanged: The gap was 26.6 in 1996 and 26.7 in 2005. Likewise, the achievement gap in NAEP fourth-grade math between Poor and Nonpoor students also did not change significantly; the gap was 25.3 in 1996 and 22.3 in 2005. A closer look at the post-NCLB socioeconomic gap trend also reveals no change. For example, the average reading achievement gap between Nonpoor and Poor students stayed about the same at grade 4 and grade 8 between 2002 and 2005. This post-NCLB trend was not different from the pre-NCLB trend, which also showed no significant changes in the gap. The average math achievement gap between Nonpoor and Poor students dropped since 2000, but the amount of recent changes after NCLB was very small: less than one point at grade 4 and less than two points at grade 8 between 2003 and 2005 (Figure 9).

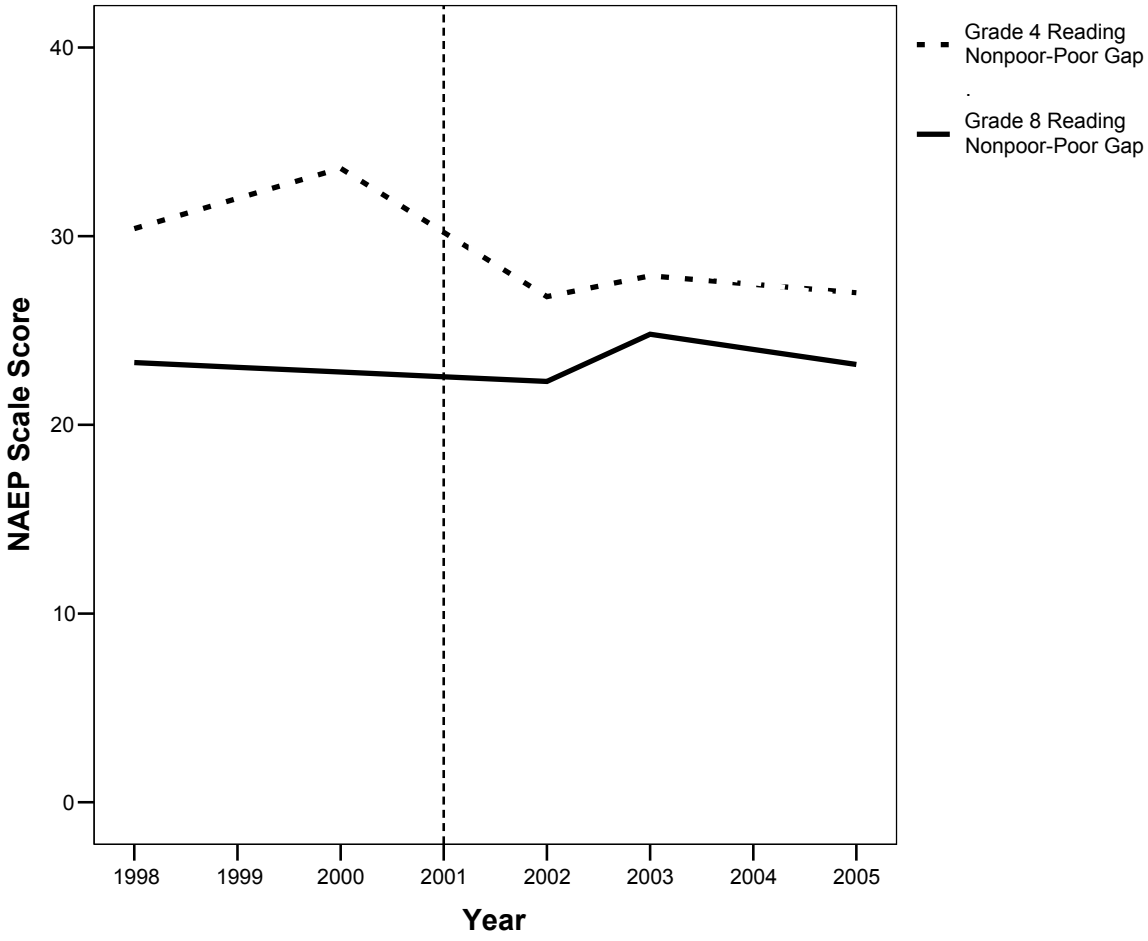

Figure 8: 1998-2005 NAEP Nonpoor-Poor Gap Trends in Grade 4 and Grade 8 Reading

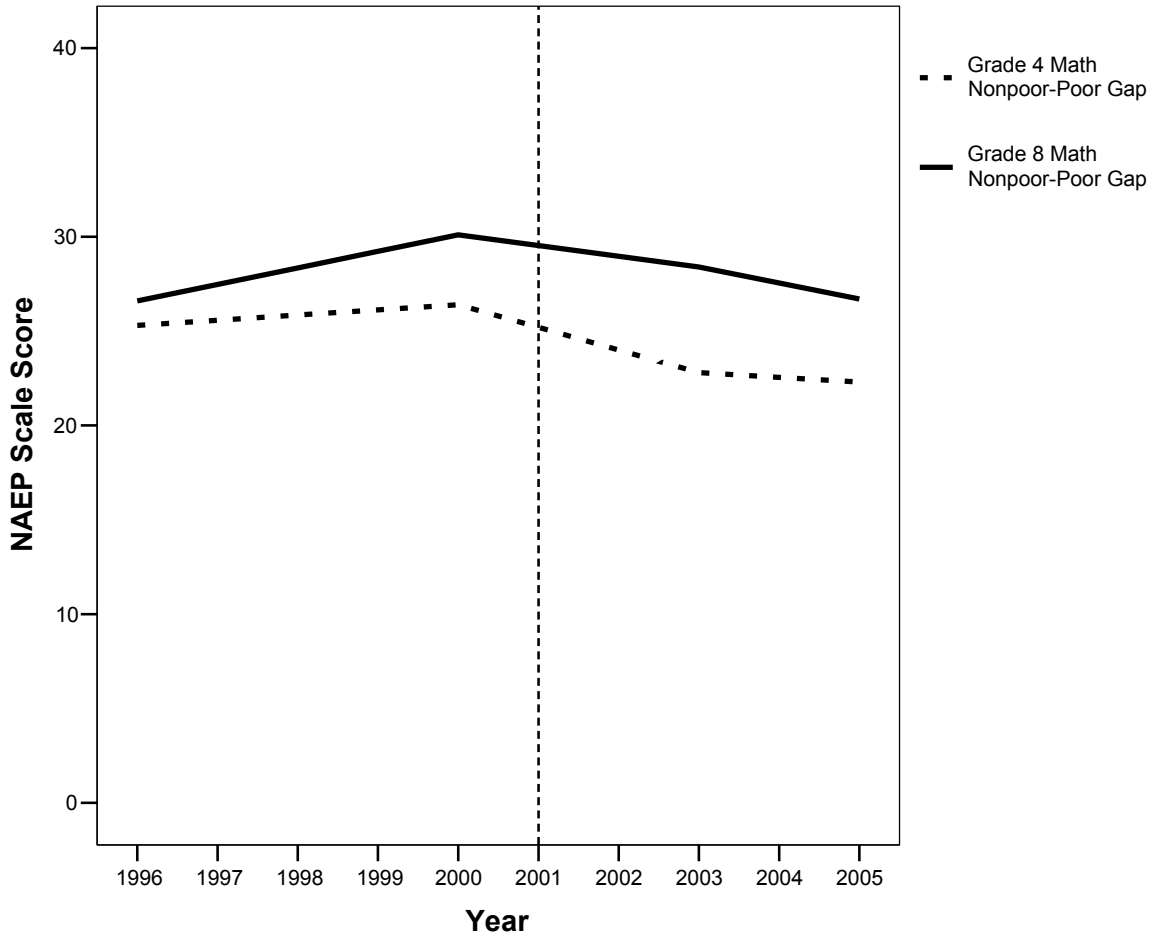

Figure 9: 1996-2005 NAEP Nonpoor-Poor Gap Trends in Grade 4 and Grade 8 Math

Pre/Post-NCLB NEAP Trends: In order to test for the statistical significance of the trends described above for each subject and grade, time-series regression analyses of national NAEP public school students' 1992-2005 reading and 1990-2005 math scale scores were conducted. The results are summarized in Table 1 for reading and in Table 2 for math; both pre-NCLB and post-NCLB growth patterns for each subgroup and their achievement gaps are classified by the significance and direction of changes. The pre-NCLB growth dimension (rows in the Tables 1 and 2) tells how the outcome measures for each group changed before NCLB: up (significantly upward trend), down (significantly downward trend); flat (no significant trend). The post-NCLB change dimension (columns in the Tables 1 and 2) tells how the pre-NCLB growth pattern changed after NCLB: increment (significant post-NCLB gain); decrement (significant post-NCLB loss); same (no significant change). Full information on the estimates of Pre-NCLB growth and Post-NCLB change parameters are presented in Tables C-1 and C-2 in Appendix C.

Table 1: National Pre-NCLB and Post-NCLB Trends in NAEP Grade 4 and Grade 8 Reading Achievement by Subgroups and their Gaps

		Post-NCLB Change		
		Increment	Same	Decrement
Pre-NCLB Growth	Up		Hispanic (8), Asian (4)	All (8), White (8), Black (8), Nonpoor (8)
	Flat		All (4), White (4), Black (4), Hispanic (4), Asian (8), Nonpoor (4), Poor, White-Black gap, White-Hispanic gap, Poverty gap	
	Down			

Note. Numbers in parenthesis refer to grades in which different growth patterns are observed. When the same growth patterns apply to both grades 4 and 8 in each subgroup or gap, no numbers are shown after the group or gap name. For the 'all' and each subgroup categories, 'up' means improvement of the average, whereas 'down' means decline of the average. For the racial and poverty gaps, 'up' signifies widening of the gap, whereas 'down' signifies narrowing of the gap.

In contrast with trends in math, the national trend in NAEP reading achievement has followed more mixed growth patterns through the 1992-2005 period (see Table 1). As already shown by graphs, the average reading achievement trend tends to be flatter than the average math achievement trend. For grade 4, there was no significant improvement at all throughout the entire period. None of the pre-NCLB growth and post-NCLB change estimates in reading (except for pre-NCLB growth among Asians) are significant. For grade 8, the average reading score improved significantly during the pre-NCLB period, but this gain has dropped after NCLB, signifying some setback in national reading progress. The pre-NCLB growth estimates in reading are significantly positive, whereas the post-NCLB change estimates are significantly negative. At the same time, the gaps among racial and socioeconomic groups in both grade 4 and grade 8 reading remained the same, and there were no significant changes in the gaps before or after NCLB.

Table 2: National Pre-NCLB and Post-NCLB Trends in NAEP Grade 4 and Grade 8 Math Achievement by Subgroups and their Gaps

		Post-NCLB Change		
		Increment	Same	Decrement
Pre-NCLB Growth	Up	Hispanic (8)	All, White, Black	
	Flat		Hispanic (4), Asian, Nonpoor, Poor, White-Black gap, White-Hispanic gap (4), Poverty gap	White-Hispanic gap (8)
	Down			

Note. Numbers in parenthesis refer to grades in which different growth patterns are observed. When the same growth patterns apply to both grades 4 and 8 in each subgroup or gap, no numbers are shown after the group or gap name. For the 'all' and each subgroup categories, 'up' means improvement of the average, whereas 'down' means decline of the average. For the racial and poverty gaps, 'up' signifies widening of the gap, whereas 'down' signifies narrowing of the gap**.**

As shown in Table 2, the results of NAEP 4th and 8th grade math trend analysis show that the national average level of NAEP math achievement improved significantly throughout the pre-NCLB period in math (except for a few subgroups). As shown by the "Pre-NCLB Growth" rows of Table 2, the national average math achievement tended to improve significantly (by about 1-2 points every year on average). However, comparison of the growth rates between the pre-NCLB period and the post-NCLB period reveals that there was no change in the rate of growth after NCLB. As shown by the "Post-NCLB Change" columns of Table 2, none of them are significant. For example, the national average 8th grade math achievement recorded a significant annual gain of .9 before NCLB, but the increment of .25 in its growth rate after NCLB was not significant. While the average math score continued to rise and reached an all time record high in 2005, there is no indication that the improvement of average math scores accelerated after NCLB across the board. The only exception to this general pattern is Hispanic 8th grade, which appears to have gained further after NCLB.

While the overall math achievement trend for average 4th and 8th grade students showed some progress, racial and socioeconomic achievement gaps remained the same throughout the 1990-2005 period (Table 2). For example, the White-Black 4th grade math gap appears to have dropped by .39 per year on average before NCLB and then further by .62 after NCLB. However, neither pre-NCLB nor post-NCLB change is large enough to be significant. Likewise, the Nonpoor-Poor 4th grade math gap remains unchanged, as both pre-NCLB growth (-.16) and post-NCLB change (-.52) are insignificant. This

suggests that all subgroups made about the same amount of achievement gains after NCLB as they did before, and that the achievement gaps did not narrow or widen significantly following the implementation of NCLB. The only exception to this general pattern is the White-Hispanic 8th grade math gap, which narrowed significantly after NCLB.

National NAEP Reading and Math Proficiency Trends

Trends in the Average Proficiency: In this section, the percent of students scoring at or above the proficient level on the NAEP are examined and the pre-NCLB trends in proficiency are compared to the post-NCLB trends. NAEP proficiency levels instead of scale scores are used. This trend analysis focuses on the percentage of students meeting or exceeding the desired NAEP performance standard, that is, students performing at or above the "Proficient" level.

The percentage of students nationally scoring at or above proficient on the NAEP in reading and math did not change significantly after NCLB (Figure 10 and 11). If we assume that the nation stays on the current trajectory, the results of trend analysis project that by 2014 only 24 to 34 percent of students would meet the reading proficiency target and about 29-64 percent of students would meet the math proficiency target.[11]

[11] Since these projections are based on the results of both grade 4 and grade 8 samples and there are often divergence of the trends between the two grades (e.g., faster growth in grade 4 than in grade 8 in math), a relatively wide range of estimates is given.

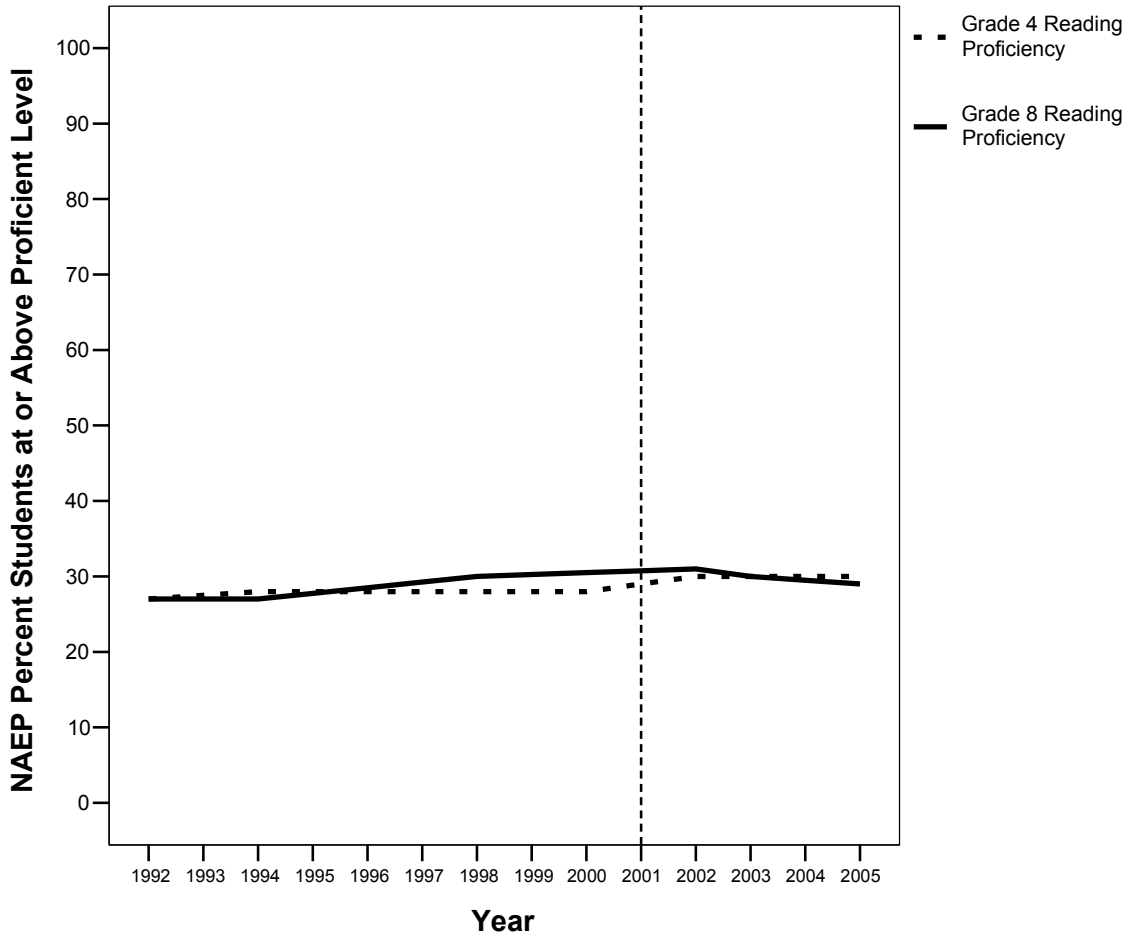

Figure 10: 1992-2005 NAEP Proficiency Rate Trends in Grade 4 and Grade 8 Reading

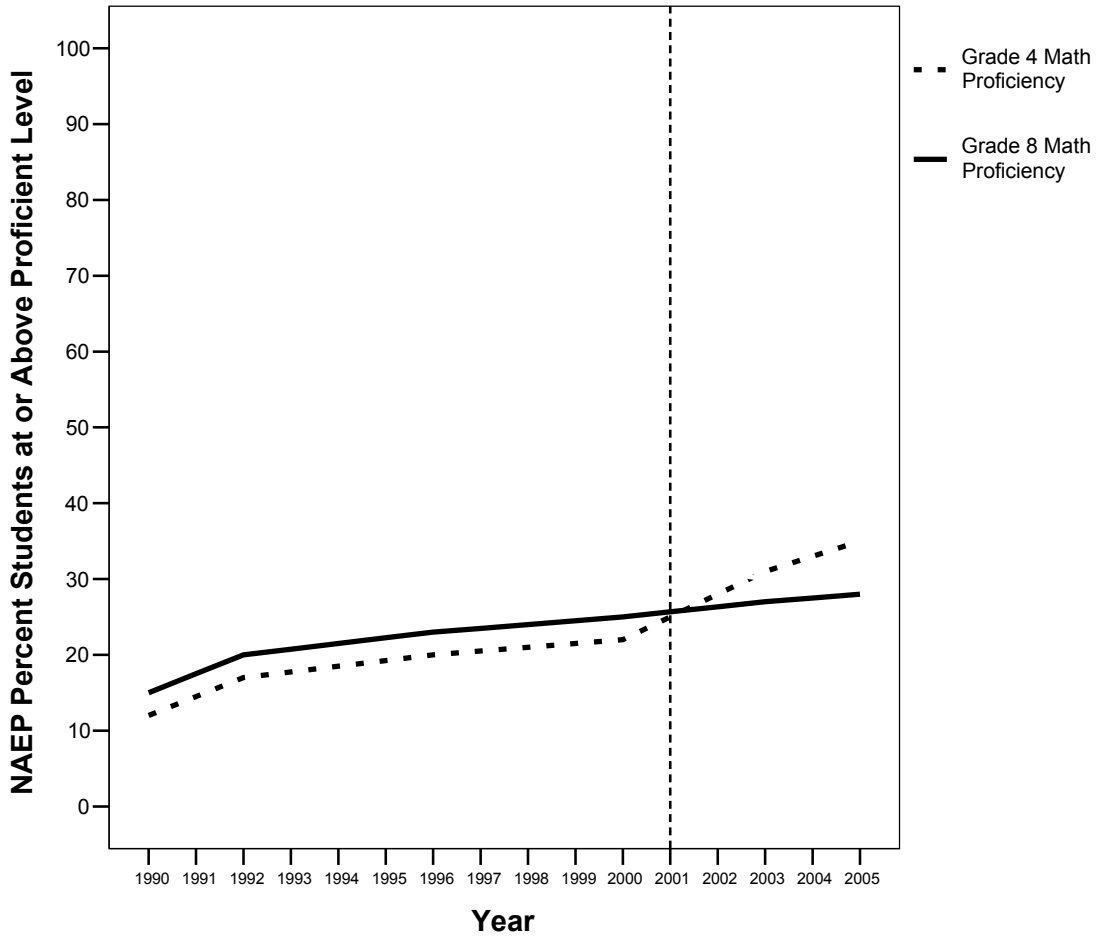

Figure 11: 1990-2005 NAEP Proficiency Rate Trends in Grade 4 and Grade 8 Math

Trends in the Proficiency Gap: If current trends in the racial and socioeconomic achievement gaps continue, substantial disparities in proficiency rates between advantaged White and disadvantaged minority groups will persist. Under the assumption that the current trajectories will continue, it is projected that by 2014 between 32 and 44 percent of Whites will reach the reading proficiency target and 40-78 percent will reach the math proficiency target. In contrast, 7 to 18 percent of Blacks will achieve proficiency in reading and 25-55 percent in math. Among Hispanics, 14 to 21 percent will achieve proficiency in reading and 32-70 percent will achieve proficiency in math. Thirty-two to thirty-eight percent of Nonpoor students will achieve proficiency in reading and 47-81 percent will reach proficiency in math, whereas only 11-16 percent of Poor students will achieve proficiency in reading and 20-76 percent in math. Obviously the feasibility of reaching the proposed goal of 100 percent proficiency raises serious concerns, particularly for disadvantaged minority students and their schools.

Pre/Post-NCLB NEAP Trends: In order to test for the statistical significance of the trends described above for each subject and grade, time-series regression analyses of national NAEP public school students' 1992-2005 reading and 1990-2005 math proficiency rates were conducted. Table C-3 and Table C-4 in Appendix C summarize the results of these statistical analyses for grade 4 and grade 8 respectively. This proficiency trend analysis shows a similar pattern of growth in the percentage of students reaching proficiency compared to the previous analysis of scale scores. The math proficiency rate has risen continuously for grades 4 and 8, except that the 8th graders' proficiency growth tends to be restricted to the White group and Nonpoor group only. For example, the pre-NCLB annual growth rate of .057 in 8th grade math proficiency for all students is significant, but the post-NCLB change of -.048 is not significant. The reading proficiency rate has been flat for grade 4 and mixed for grade 8 (significant earlier gains followed by losses after NCLB). Moreover, the relative gaps in the proficiency rate among racial and socioeconomic groups also remained largely unchanged.

There was a period when the racial achievement gap narrowed substantially with significant academic progress of Blacks and Hispanics. Prior research based on the long-term trend NAEP showed that the racial achievement gaps narrowed at the basic skills level in the 1970s and early 1980s but grew at the advanced skills level in the late 1980s and the 1990s (Lee, 2002). Even if there were significant reductions of the achievement gaps in certain areas after NCLB, they may be viewed as relatively much small when compared to the magnitude of the past decreases. The overall Black-White and Hispanic-White achievement gaps remained substantially large at the high achievement level that current NAEP proficiency standard signifies.

PART 3: STATE ACHIEVEMENT TRENDS IN NAEP

Notwithstanding the aggregate national NAEP trends, there are substantial variations among states in growth patterns on the NAEP state assessment. This PART examines changes in NAEP scores using data from the NAEP state assessments. In addition to the national NEAP, which is based on a nationally representative sample of students, the state assessments are based on representative sample of public school students selected from participating states. The first section presents the results of NAEP scale score analysis and the second section presents the results of the proficiency rate analysis for individual states.

State NAEP Reading and Math Scale Score Trends

Trends in the Average Achievement: The NAEP state assessment has provided information on state-by-state reading and math achievement for grade 4 and grade 8 since 1990. Using data from the NAEP state assessment, a baseline level of performance can be determined for each state, and it can be used to compare states to each other. There is considerable variability among states in baseline scores. For example, the states' baseline status of grade 4 math average score as of 1992 varies from 201 in Mississippi to 231 in Maine. States also vary in the amount of growth rate they achieve over time, with some states making greater progress towards improving NAEP scores than others. For example, the states' annual growth rate for grade 4 math average score before NCLB varies significantly from .4 in Maine to 1.6 in North Carolina. Over time, achievement gains in states that are relatively low-performing at the baseline tend to be less than in states with higher baseline scores ($r = -.63$ between initial status and pre-NCLB growth rate in grade 4 math). On the other hand, the post-NCLB change in grade 4 math annual growth rate does not vary significantly among the states, ranging from 1.5 in North Carolina to 2 in New Mexico. While most states continued their pre-NCLB growth pattern after NCLB, states that made relatively larger achievement gains before NCLB tend to accelerate at a slower rate after NCLB ($r = -.58$ between pre-NCLB growth and post-NCLB change in grade 4 math).

Table 3 and Table 4 classify states based on the results of HLM trend analyses of NAEP 4th grade and 8th grade state average scores in reading and math respectively. As with the national trend, the growth trajectory was divided into pre-NCLB and post-NCLB time periods. In order to test for the statistical significance of the trends, HLM growth modeling analyses of state NAEP reading and math scale scores were conducted. Pre-NCLB growth dimension (rows in the Tables 1 and 2) tells which states changed in which directions before NCLB: up (significantly upward trend), down (significantly downward trend); flat (no significant trend). Post-NCLB change dimension (columns in the Tables 1 and 2) tells whether and how their pre-NCLB growth pattern changed after NCLB: increment (significant post-NCLB gain); decrement (significant post-NCLB loss); same (no significant change). For full information on the estimates of both Pre-

NCLB growth and Post-NCLB change parameters for racial and socioeconomic subgroups as well as the average students, see Tables C-5 and C-6 in Appendix C.[12]

[12] For racial minority groups, the number of states is less than 50 since the NAEP test results for Asians, Blacks and/or Hispanics in certain states are not available due to insufficient sample size of the groups for reliable estimation. For each minority group, states that are not included in the Tables are as follows:
(1) Asian
Alabama, Arizona (except grade 4 reading and math), Arkansas, Idaho, Indiana, Iowa (except grade 4 reading), Kentucky, Louisiana, Maine, Michigan (except grade 4 math), Mississippi, Missouri, Montana, Nebraska, New Hampshire, New Mexico, North Dakota, Ohio, Oklahoma (except grade 4 math), Pennsylvania (except grade 4 reading, grade 8 reading and math), South Carolina, South Dakota, Tennessee, Vermont, West Virginia, and Wyoming.
(2) Black
Hawaii (except grade 4 reading and math, grade 8 reading), Idaho, Maine, Montana, New Hampshire, North Dakota, South Dakota, Utah,
Vermont, and Wyoming
(3) Hispanic
Alabama, Kentucky, Louisiana, Maine, Mississippi, Missouri (except grade 4 reading and math, grade 8 reading), Montana (except grade 4 reading and math), New Hampshire (except grade 4 reading and math), North Dakota, South Carolina (except grade 4 reading and math, grade 8 math), South Dakota (except grade 4 math), Tennessee (except grade 4 reading and math), Vermont, and West Virginia

Table 3: Classification of States in Pre-NCLB and Post-NCLB Trends of NAEP Grade 4 and Grade 8 Reading Average Achievement

		Post-NCLB Change		
		Increment	Same	Decrement
Pre-NCLB Growth	Up		CO(4), DE(4), FL(4), MD(4), MO(8), NY(4)	DE(8)
	Flat		AL, AK, AZ, AR, CA, CO (8), CT, FL(8), GA, HI, ID, IL, IN, IA, KS, KY, LA, ME, MD(8), MA, MI, MN, MS, MO(4), MT, NE, NV, NH, NJ, NM, NY(8), NC, ND, OH, OK, OR, PA, RI, SC, SD, TN, TX, UT, VT, VA,WA, WV, WI, WY	
	Down			

Note. Numbers in parenthesis refer to grades in which different growth patterns are observed. When the same growth patterns apply to both grades 4 and 8 in each state, no numbers are shown after state code. 'Up' means improvement of the average, whereas 'Down' means decline of the average.

Table 4: Classification of States in Pre-NCLB and Post-NCLB Trends of NAEP Grade 4 and Grade 8 Math Average Achievement

		Post-NCLB Change		
		Increment	Same	Decrement
Pre-NCLB Growth	Up	AL(4), AZ(4), AR(4), CA(4), CO(4), CT(4), DE(4), FL(4), GA(4), HI(4), ID(4), IN(4), KS(4), KY(4), LA(4), MD(4), MA(4), MI(4), MN(4), MS(4), MO(4), NH(4), NJ(4), NY(4),NC(4), OH(4), OK(4), OR(4), PA(4), RI(4), SC(4), TN(4), TX(4), UT(4), VT(4), VA(4), WA(4), WV(4), WY(4)	AL(8), AZ(8), AR(8), CA(8), CO(8), CT(8), DE(8), FL(8), GA(8), HI(8), ID(8), IL(8), IN(8), KY(8), LA(8), MD(8), MA(8), MI(8), MN(8), MS(8), NH(8), NJ(8), NY(8), NC(8), OH(8), OR(8), PA(8), RI(8), SC(8), TX(8), VA(8), WV(8), WI(8), WY(8)	
	Flat	AK(4), IL(4), IA(4), ME(4), MT(4), NE(4), NV(4), NM(4), ND(4), SD(4), WI(4)	AK(8), IA(8), KS(8), ME(8), MO(8), MT(8), NE(8), NV(8), NM(8), ND(8), OK(8), SD(8), TN(8), UT(8), VT(8), WA(8)	
	Down			

Note. Numbers in parenthesis refer to grades in which different growth patterns are observed. When the same growth patterns apply to both grades 4 and 8 in each state, no numbers are shown after state code. 'Up' means improvement of the average, whereas 'Down' means decline of the average.

In reading, most states did not make progress in improving average scores at both grades levels either before or after NCLB. In math, many states made significant gains at both grades before NCLB, and they continued the same rate of progress (grade 8) or accelerated their progress (grade 4) after NCLB. For example, the annual growth rate of NAEP grade 4 math average score was, on average across all states, about 1 point (M = .96) during the pre-NCLB period. This pre-NCLB growth rate varied significantly among states, ranging from .4 to 1.6 (SD = .35). Among 50 states, 39 states showed a significant

upward trend, whereas 11 states were flat. After NCLB, this pre-NCLB growth rate increased by 1.7 points (M = 1.70). The post-NCLB increment of annual math gain score in all 50 states was significant, ranging from 1.5 to 2 (SD = .14).

Trends in the Achievement Gap: Further, there are also variations among states in racial achievement gap trends. Despite substantial variations in the initial status of the gap at the baseline, the White-Black gap tends to remain flat throughout the period for most states. For example, the states' baseline status of grade 4 math White-Black gap as of 1992 varies from 15 points in West Virginia to 42 points in Michigan. However, states do not vary much in their growth rate, as most states made little or no progress in narrowing the gap. For example, the states' annual growth rate of grade 4 math White-Black gap before NCLB varies from -.01 in West Virginia to -.83 in Minnesota. Similar patterns continued after NCLB.

Table 5 and Table 6 classify states based on the results of HLM trend analyses of NAEP 4th grade and 8th grade state average White-Black gaps in reading and math respectively. In grade 8 reading and math, none of the states changed the White-Black gaps in either the pre-NCLB or post-NCLB period. While there were some variations among states in the amount of changes, they were not significant. For example, the amount of pre-NCLB annual change in Black–White eighth-grade math test score gaps (M = .13 SD = .33), and in Hispanic–White eighth-grade math test score gaps (M = .20, SD = .71) were all insignificant. States also did not make significant changes in the gap between Poor and Nonpoor students throughout the period; M = .50, SD = .72 for pre-NCLB growth; M = -1.08, SD = 1.36 for post-NCLB change. Consequently, racial and socioeconomic achievement gaps did not significantly change after NCLB in most states. For full information on the estimates of both Pre-NCLB growth and Post-NCLB change for racial and socioeconomic achievement gaps, go to Tables C-5 and C-6 in Appendix C.

Table 5: Classification of States in Pre-NCLB and Post-NCLB Trends of NAEP Grade 4 and Grade 8 Reading White-Black Gap

		Post-NCLB Change		
		Increment	Same	Decrement
Pre-NCLB Growth	Up			
	Flat		AL, AK, AZ, AR, CA, CO, CT, DE, FL, GA, HI, IL, IN, IA, KS, KY, LA, MD, MA, MI, MN, MS, MO, NE, NV, NJ, NM, NY, NC, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WV, WI	
	Down			

Note. Numbers in parenthesis refer to grades in which different growth patterns are observed. When the same growth patterns apply to both grades 4 and 8 in each state, no numbers are shown after state code. 'Up' signifies widening of the gap, whereas 'Down' signifies narrowing of the gap.

Table 6: Classification of States in Pre-NCLB and Post-NCLB Trends of NAEP Grade 4 and Grade 8 Math White-Black Gap

		Post-NCLB Change		
		Increment	Same	Decrement
Pre-NCLB Growth	Up			
	Flat		AL, AK, AZ, AR, CA, CO, CT, DE, FL, GA, HI (4), IL, IN, IA, KS, KY, LA, MD, MA, MI, MN(8) MS, MO, NE, NV, NJ, NM, NY, NC, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WV, WI	
	Down		MN(4)	

Note. Numbers in parenthesis refer to grades in which different growth patterns are observed. When the same growth patterns apply to both grades 4 and 8 in each state, no numbers are shown after state code. 'Up' signifies widening of the gap, whereas 'Down' signifies narrowing of the gap.

Do the results of this state achievement trend analysis give the same or different information from the national achievement trend analysis as reported in Part II? By and large, the results of combining state-level data also imply divergent trends between reading and math. In reading, only a handful of states made significant gains before NCLB, and none accelerated its growth after NCLB. In math, many states made significant gains on average (except for some racial and socioeconomic groups) throughout the pre-NCLB period at both grade 4 and grade 8. However, post-NCLB progress towards improving math achievement was mixed. Fourth graders' math achievement accelerated since NCLB, while eighth graders' math achievement stayed the same course of growth. The results of this state-level analysis, showing a significant, post-NCLB change in the state average grade 4 math achievement, contrasts with the corresponding national-level aggregate pattern of insignificant post-NCLB change. It needs to be noted that significant post-NCLB improvement of grade 4 math achievement in many states occurred mostly between 2000 and 2003 but not between 2003 and 2005; the temporary increase was followed by a return to the pre-reform growth rate. Finally, the results of national and state analyses converge with regard to equity, in that the achievement gaps among racial and socioeconomic groups in both reading and math remained largely unchanged throughout the entire period.

State NAEP Reading and Math Proficiency Trends

Trends in the Average Proficiency: Similar patterns are found from the results of HLM analyses that investigate the reading and math trends in proficiency rates using all states' NAEP assessment data throughout the 1990-2005 period and includes both pre-NCLB and post-NCLB time blocks. In both reading and math, there were significant gains made by all racial and socioeconomic groups throughout the period at grade 4 and 8. However, progress was mixed after NCLB. The trend in fourth graders' math achievement has accelerated since NCLB. On the other hand, there were also some significant setbacks after NCLB, including the deceleration of 8th grade reading.

Trends in the Proficiency Gap: The gaps in reading and math proficiency rates among racial and socioeconomic groups remained largely unchanged throughout the period. Exceptions to this pattern were the White-Black gap in 4th grade math and the Nonpoor-Poor gap in 4th and 8th grade reading, both of which narrowed significantly throughout the period. The Nonpoor-Poor gap in grades 4 and 8 reading grew since NCLB while it went down in grade 4 math. All other gaps remained the same and the earlier gap pattern perpetuated since NCLB.

Effects of State Accountability Policies on the NAEP Reading and Math Achievement Trends

This study tests the hypothesis that the first generation accountability states that had high-stakes testing and a strong accountability system in the 1990s would have had greater academic improvement before NCLB, whereas the second generation accountability states that lacked such a system in the 1990s would make greater progress after NCLB. To test the hypothesis, this study uses the measure of state accountability constructed by Lee and Wong (2004) (see Appendix B for description of variable). Based on this accountability policy score, 50 states were also classified into three groups: strong accountability systems (13 states in the top quartile), those with moderate accountability systems (25 states in the middle half), and states with weak accountability systems (12 states in the bottom quartile).[13] Although most weak accountability states also had state assessments, and some even had report cards for schools, none of them provided direct incentives to schools in the form of performance ratings, rewards, assistance, and/or sanctions. This weak accountability group represents the second generation accountability states. In contrast, most strong accountability states turned out to have these key elements of accountability policy in place, and this group represents the first generation accountability states.

[13] States with strong accountability systems include Alabama, Florida, Illinois, Indiana, Kentucky, Louisiana, Maryland, New Jersey, New Mexico, New York, North Carolina, and Texas. In contrast, states with weak accountability systems include Alaska, Arkansas, Colorado, Delaware, Idaho, Iowa, Maine, Massachusetts, Montana, Nebraska, New Hampshire, North Dakota, and Wyoming. Strong accountability states are more likely to be the first-generation accountability states, whereas weak accountability are more likely to be the second-generation accountability states which did not have statewide high-stakes testing and accountability systems until NCLB.

If we were to find significantly positive effect of this state accountability variable on pre-NCLB growth, but at the same time significantly negative effect on post-NCLB change, it would support the above hypothesis. Table C-7 and Table C-8 in Appendix C summarize the results of HLM analysis on the relationship of state accountability with pre- and post-NCLB reading and math achievement trends at grade 4 and grade 8 respectively.

The Effects of State Accountability on the Average Achievement: Some individual states made relatively larger academic progress than other states, but this progress does not appear to be systematically related to the kinds of state reform variables that might support the hypothesis of long-term test-driven external accountability policy. When pre-NCLB and post-NCLB achievement trends appear to favor test-driven accountability, this phenomenon seems to partly reflect an artifact of regression to the mean; the first generation states were performing low at the baseline and made relatively larger math achievement gains prior to NCLB than the second generation states. Further, the findings imply that NCLB did not work yet as intended to transfer the alleged effects of a test-driven external accountability system to all states.

With consistently insignificant effects in reading, it appears that state accountability policy contributes very little to the interstate variation in the NAEP reading trend, whether it concerns pre-NCLB growth or post-NCLB change. An exception is found in the grade 4 reading trend for Whites only. In contrast, it appears that the state accountability variable contributes partly to the pre-NCLB growth in math but not to the post-NCLB change. In other words, the earlier accountability policy effect on math achievement among the first generation states, if any, fails to have transferred to the second generation states as a result of NCLB as shown by insignificant policy effects on post-NCLB change.

Figure 12 shows that the state average NAEP grade 4 math achievement gain prior to NCLB was relatively larger in strong accountability states than in weak accountability states. For example, the HLM estimate of pre-NCLB annual growth rate for Whites was 1.05 in the strong accountability states and .87 in the weak accountability states. This difference in annual gain translates into cumulative gains of 10.5 and 8.7 for each group over the past 10 years prior to NCLB (1992-2001). Although the difference of 1.8 was statistically significant, it may not be of practical import. After adjustment for initial status differences between strong and weak accountability states, significantly positive state accountability policy effect observed prior to NCLB is limited to only Whites in grade 4 math and only Whites and Hispanics in grade 8 math (Table C-7 and C-8).

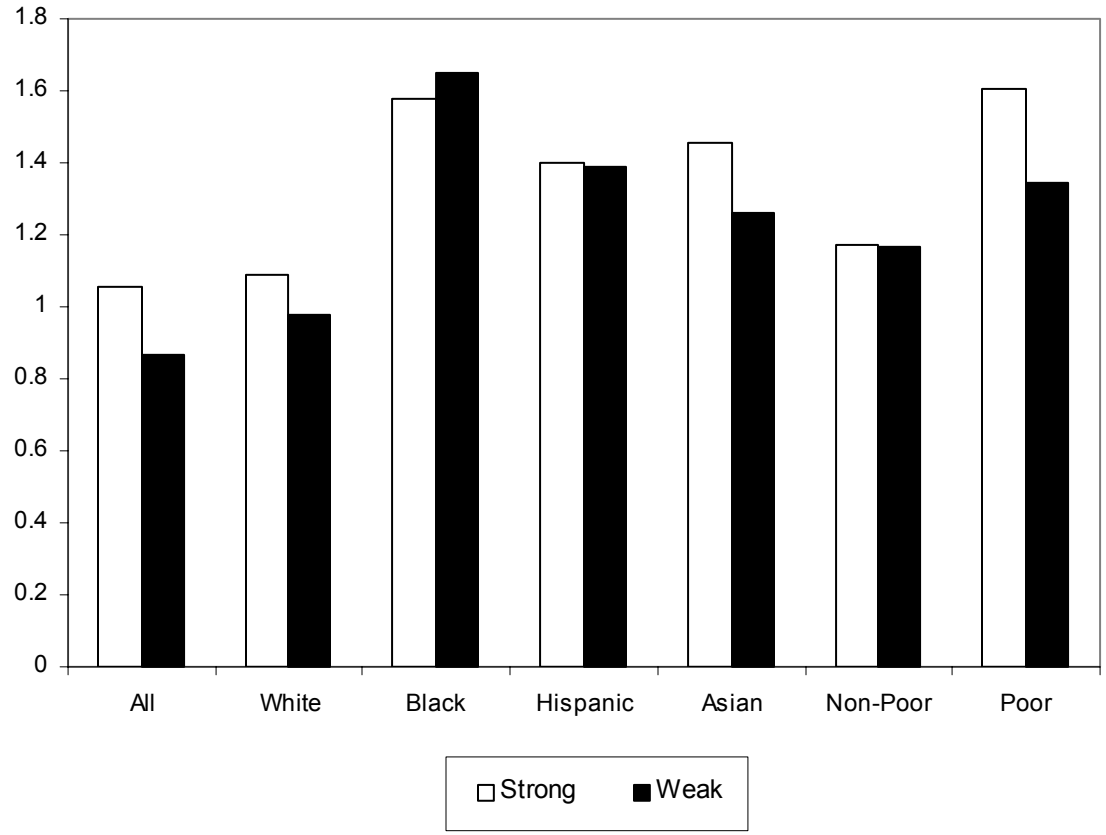

Figure 12: Strong vs. Weak Accountability States' Average pre-NCLB Annual Growth Rates in NAEP Grade 4 Math Achievement by Subgroup

On the other hand, there were no clear indications from the analysis of post-NCLB change patterns (2002-05) that the math achievement of 4th grade students improved more in the second generation states than in the first generation states. Figure 13 compares post-NCLB change in the states' average grade 4 math annual growth rates between strong and weak accountability states. For example, the HLM estimate of post-NCLB change in grade 4 math growth rate for Whites was 1.55 in the strong accountability states and 1.56 in the weak accountability states. This difference is not significant, and thus it does not give support for the claim that the states without test-driven external accountability policy before NCLB should benefit more from NCLB than the states with preexisting accountability. While some significant policy effects were observed among Hispanic and Poor students, the relationships turned out to be very tenuous once we take into account for differences in initial status and pre-NCLB growth rate (Table C-7 and Table C-8).

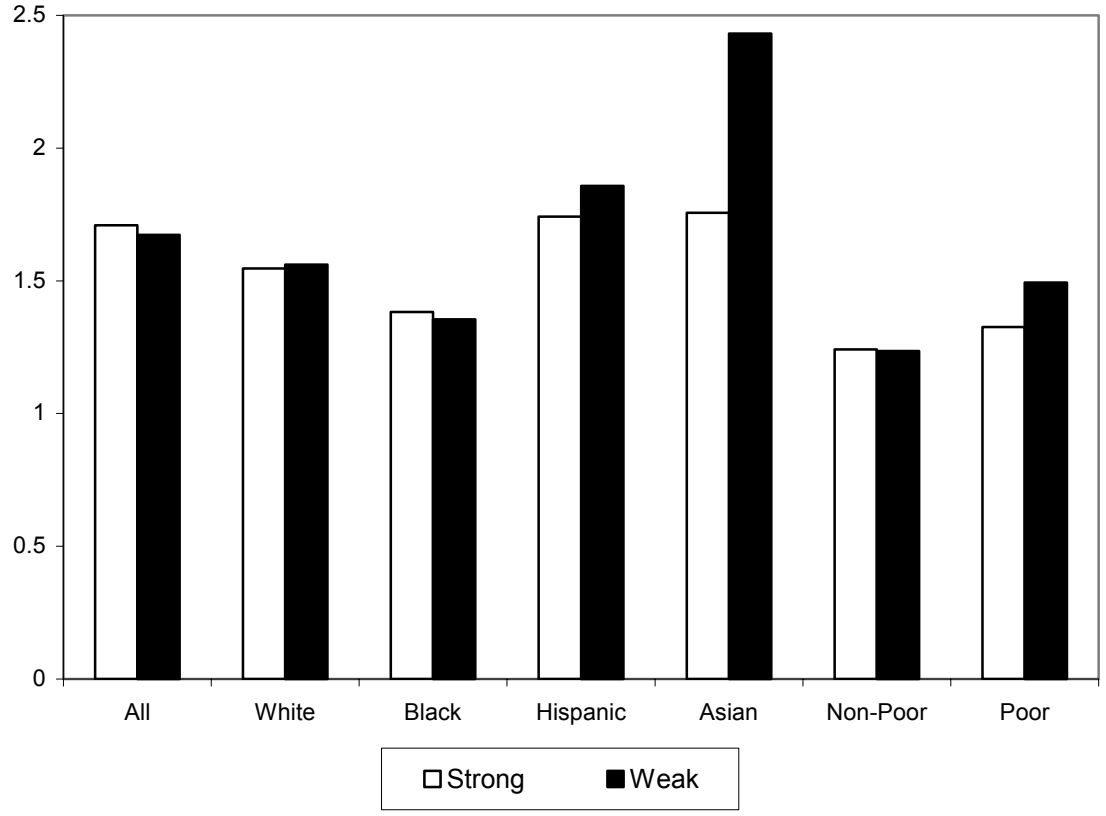

Figure 13: Strong vs. Weak Accountability States' Average post-NCLB Change to Annual Growth Rates in NAEP Grade 4 Math Achievement by Subgroup

The Effects of State Accountability on the Achievement Gap: Further, HLM analyses were also conducted to test the effects of accountability policies on racial and socioeconomic achievement gaps. The results of HLM analyses suggest that although the strong accountability states such as Texas with initially larger achievement gaps appear to have narrowed some of the gaps more than their weak accountability counterparts before NCLB, there is no significant difference between the two groups of states once their initial difference was considered. Further, there is no indication that the gaps narrowed more or less in one group of states than the other after NCLB.

By and large, state accountability was not significantly related to pre-NCLB growth and post-NCLB changes in the racial and socioeconomic gaps. In both reading and math, few states changed the gaps significantly over the entire period, and there were no systematic differences between strong accountability states and weak accountability states in terms of changes in achievement gaps for Blacks and Hispanics as well as for Poor students. For example, Texas, one of strong accountability states, appears to have made some progress between 1990/92 and 2005 in narrowing the White-Black gap (1 point increase in grade 4 reading; 5 point decrease in grade 4 math; 7 point decrease in grade 8 math) and the White-Hispanic gap (1 point decrease in grade 4 reading; 3 point

decrease in grade 4 math; 4 point decrease in grade 8 math). Progress is also seen in the state's narrowing the Nonpoor-Poor gap between 1996/98 and 2005 (5 point decrease in grade 4 reading and math, 2 point decrease in grade 8 reading; 5 point decrease in grade 8 math). Considering initially large gaps at the baseline (e.g., 38 points for White-Black gap and 28 points for White-Hispanic gap in 1990 grade 8 math), however, these reductions only account for 4-20 percent of the initial gaps.

PART 4: DISCREPANCIES BETWEEN NAEP AND STATE ASSESSMENT RESULTS

NCLB requires each state to develop a test-based accountability system to monitor the performance of schools and districts. Each state administers its own assessments and establishes performance targets that students must meet.[14] Although NCLB establishes state assessments as the basis for NCLB accountability, NAEP can play a confirmatory role as an independent assessment to validate the state test results. Using two measures of states' academic performance (states' own assessments and the NAEP state assessments), the first section compares the percentage of students meeting or exceeding the proficiency standard in reading and math set by each state with the percentage of students meeting or exceeding the NAEP proficiency standard. Separate results are reported for racial and socioeconomic subgroups. Secondly, the role of state accountability policy in fostering improvements in student achievement is explored by examining variations among states in the patterns of discrepancies between NAEP and state assessment in the average proficiency and the achievement gap. Finally, further comparison is made between NAEP and state assessment with regard to post-NCLB academic progress as measured by the state average proficiency gains.

NAEP vs. State Assessment Results on the Average Proficiency and the Gap

The percentages of students meeting or exceeding the proficiency standard in both reading and math were, on average, twice as large, and in some cases, even larger, on state assessments than on the NAEP. This implies that for most states, NAEP performance standards are more challenging than are the states' own (see Table B-1 in Appendix B for a measure of the discrepancies between NAEP and state assessments in reading and math proficiency for each state). Figures 14 and 15 illustrate the discrepancies between NAEP scores and performance on state assessments in grade 4 reading and math respectively. There were discrepancies between the NEAP and state assessments for every racial group; the discrepancy tends to be especially large for Blacks (about 4 times larger) and Hispanics (about 3 times larger) in comparison with Whites and Asians (about 2 times larger). The discrepancies also existed for economic subgroups: Poor (about 3 times) and Nonpoor (about 2 times). This suggests that the discrepancies between NAEP and state assessment may have been larger for disadvantaged and minority groups than for advantaged and White groups. These uneven patterns of the discrepancies may result because the disadvantaged and minority groups include more low-achieving students who could have passed state standard, but not the more rigorous NAEP standard.

---

[14] While many states adopted achievement levels that are very similar to NAEP levels, the labels of achievement levels vary among states. For comparison of the assessment results related to achievement levels, this study only examines the level of achievement defined by the states as meeting desired performance standards under NCLB. On NAEP, student achievement at or above "Proficient" is treated as meeting or exceeding the national standard.

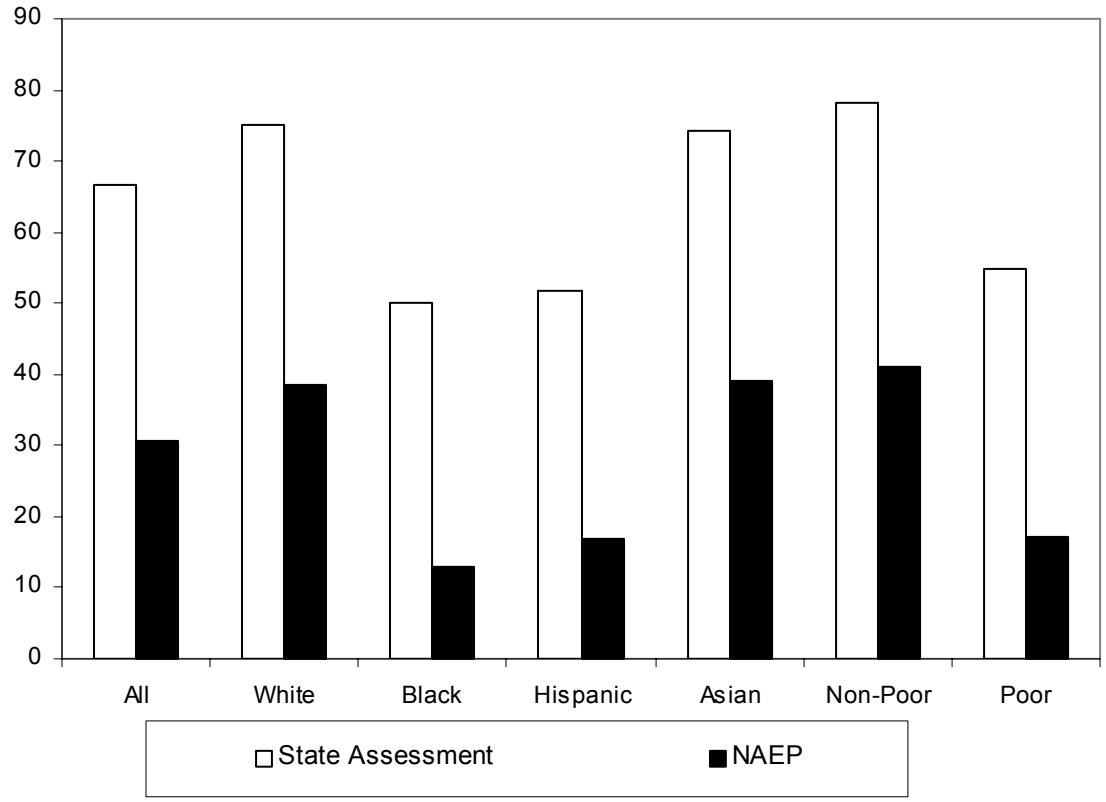

Figure 14: Percentages of Students by Subgroup Meeting or Exceeding the Proficiency Standard in Grade 4 Reading on State Assessment vs. NAEP

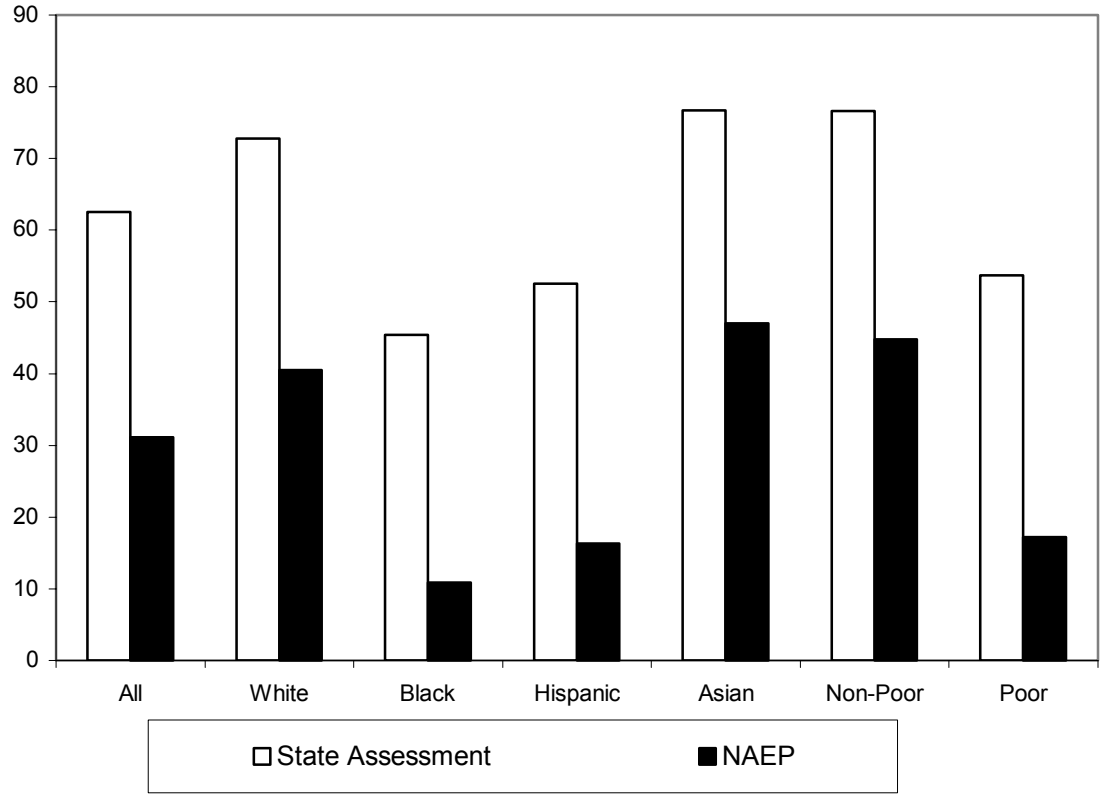

Figure 15: Percentages of Students by Subgroup Meeting or Exceeding the Proficiency Standard in Grade 4 Math on State Assessment vs. NAEP

Table 7 summarizes the discrepancies between NAEP and state assessments across states for all students and each subgroup by subject and grade. The discrepancy between the two assessments is a ratio of the state assessment-based estimate of proficiency rate to the NAEP-based estimate of proficiency rate. The more this ratio departs from the value of one, the greater the discrepancies between the two assessments. When the ratio exceeds 1, it suggests that state standards are lower than the NEAP standards, whereas a ratio below 1 suggests that state standards are relatively higher than the NAEP standards.[15] In all cases, the ratio is larger than 1, suggesting that state standards are lower than NAEP standards. The discrepancies between NAEP and state assessment results are the largest for Black, Hispanic and Poor students and the smallest for White and Nonpoor students. These findings are consistent across grades and in both reading and math. This suggests that Blacks, Hispanic and Poor students are less likely to meet the proficiency standard than White and Nonpoor students, and the proficiency gap tends to be larger with the NAEP standard than with the state standard.

[15] For the proficiency gap among racial and socioeconomic groups, a ratio greater than 1 implies state overestimation of the gap, relative to NAEP, while a ratio less than 1 implies state underestimation of the gap.

Table 7: Discrepancies between NAEP and State Assessment Results in Grade 4 and Grade 8 Reading and Math (N = 43 states)

	Ratio of state-assessment proficiency rate to NAEP proficiency rate			
	Grade 4		Grade 8	
	Reading	Math	Reading	Math
All	M= 2.25, SD=. 69	M=2.10, SD=. 71	M=2.02, SD=. 61	M=1.95, SD=. 81
White	M=1.98, SD=. 45	M=1.85, SD=. 47	M=1.86, SD=. 45	M=1.80, SD=. 57
Black	M=4.12, SD=2.12	M=4.66, SD=2.19	M=3.61, SD=1.73	M=4.47, SD=2.91
Hispanic	M=3.29, SD=1.41	M=3.64, SD=1.73	M=3.08, SD=1.27	M=3.30, SD=1.55
Asian	M=2.05, SD=. 67	M=1.75, SD=. 51	M=1.83, SD=. 55	M=1.72, SD=. 55
Nonpoor	M=1.93, SD=. 35	M=1.73, SD=. 35	M=1.87, SD=. 42	M=1.80, SD=. 44
Poor	M=3.33, SD=1.38	M=3.37, SD=1.45	M=2.95, SD=1.43	M=3.31, SD=1.91
White-Black gap	M=. 55, SD=. 17	M=. 45, SD=. 16	M=. 60, SD=. 20	M=. 53, SD=. 23
White-Hispanic gap	M=. 65, SD=. 19	M=. 57, SD=. 17	M=. 69, SD=. 21	M=. 62, SD=. 19
Poverty gap	M=. 63, SD=. 17	M=. 57, SD=. 15	M=. 73, SD=. 23	M=. 63, SD=. 21

Note. M is the average ratio of state assessment-based proficiency rate to NAEP-based proficiency rate, and SD is the standard deviation of the ratio across states and years. For the racial gap, an odds ratio was calculated by dividing the ratio of White proficiency rate to Black or Hispanic proficiency rate based on state assessment by its corresponding ratio based on NAEP. Likewise, an odds ratio was calculated for the poverty gap by dividing the ratio of Nonpoor proficiency rate to Poor proficiency rate based on state assessment by its corresponding ratio based on NAEP.

Compared to the NAEP, state assessments tend to underestimate the racial and socioeconomic achievement gap. This finding is related to uneven patterns of NAEP vs. state assessment discrepancies in proficiency rates for different racial and socioeconomic groups. As shown in the bottom of Table 7, the estimate of the achievement gap between Black and White students obtained from state assessments is half the Black-White achievement gap estimated by NAEP. For example, the White-Black gap in grade 4 math based on the state assessment was 1.8; the percentage of students meeting or exceeding

the proficiency standard was 1.8 times greater for Whites than for Blacks. In contrast, the corresponding White-Black gap based on NAEP was 4.3; the percentage of students scoring at or above Proficient was 4.3 times greater for Whites than for Blacks. Consequently, the estimate of White-Black proficiency gap based on state assessment was only half of the gap estimate based on NAEP (M = .45 for grade 4 math White-Black gap in Table 7). Likewise, the estimate of the achievement gap between Hispanic and White students obtained from state assessments is two-thirds of the estimate of the Hispanic-White achievement gap estimated by NAEP. The same pattern of discrepancy is found for the Nonpoor-Poor achievement gap.

Effects of State Accountability on the Divergence of NAEP and State Assessment Results

While the aforementioned findings reflect typical nationwide patterns, there are interstate variations in the discrepancies between NAEP proficiency standards and state proficiency standards (see standard deviations in Table 7). One factor that may explain these observed variations among states is the degree to which consequences (rewards or sanctions) are attached to state test results for schools and students. This study hypothesizes that the states that have high-stakes testing and a strong accountability system would exert greater pressure for schools and students to improve their achievement on the state test than states without high-stakes accountability systems. High stakes accountability results in the possible inflation of the number of students reaching the proficiency level and thus a deflation of the achievement gap. The state education agency itself is also likely to water down its own performance standards in anticipation of massive failure. To test the hypothesis, this study uses the measure of state accountability constructed by Lee and Wong (2004) (see Appendix B for description of the variable).

Correlation analysis supports the hypothesis that proficiency levels in states with high-stakes testing and accountability systems are inflated. We find a positive relationship between the level of state accountability and the size of NAEP-state assessment discrepancies. The discrepancies between state proficiency levels and NEAP proficiency levels are particularly large in math. The results are presented in Figure 16 and in Table C-9 in Appendix C, which summarizes the results of correlation analysis by grade and subject.

The results indicate that the higher the stakes attached to state assessments, the lower the states' own performance standards relative to NAEP standards. Figure 14 illustrates this relationship among 50 states by displaying the level of state accountability (horizontal axis) and the size of discrepancy between NAEP and state assessments in grade 8 math proficiency rate (vertical axis): the correlation between two variables is significantly positive (r = .36). For example, the stakes for failing to meet the state's performance targets are higher in Kentucky than in Maine. In 2003, 31% of students were proficient in grade 8 math on Kentucky's state test versus 24% that were proficient on the NAEP. As a result, the discrepancy between the NAEP and state assessment is 1.3 (the ratio of 31 to 24) in Kentucky. This suggests that the performance standards for the

Kentucky Core Content Test (KCCT) have been set at relatively lower levels than the standards for NAEP. In contrast, the performance standards for the Maine Education Assessment (MEA) have been set at relatively higher levels than the standards for NAEP. In 2003, 18% of students met or exceeded the standard in grade 8 math on the Maine state test, whereas 29% of students scored proficient on the NAEP. This results in the discrepancy of 0.6 (ratio of 18 to 29 in Maine).

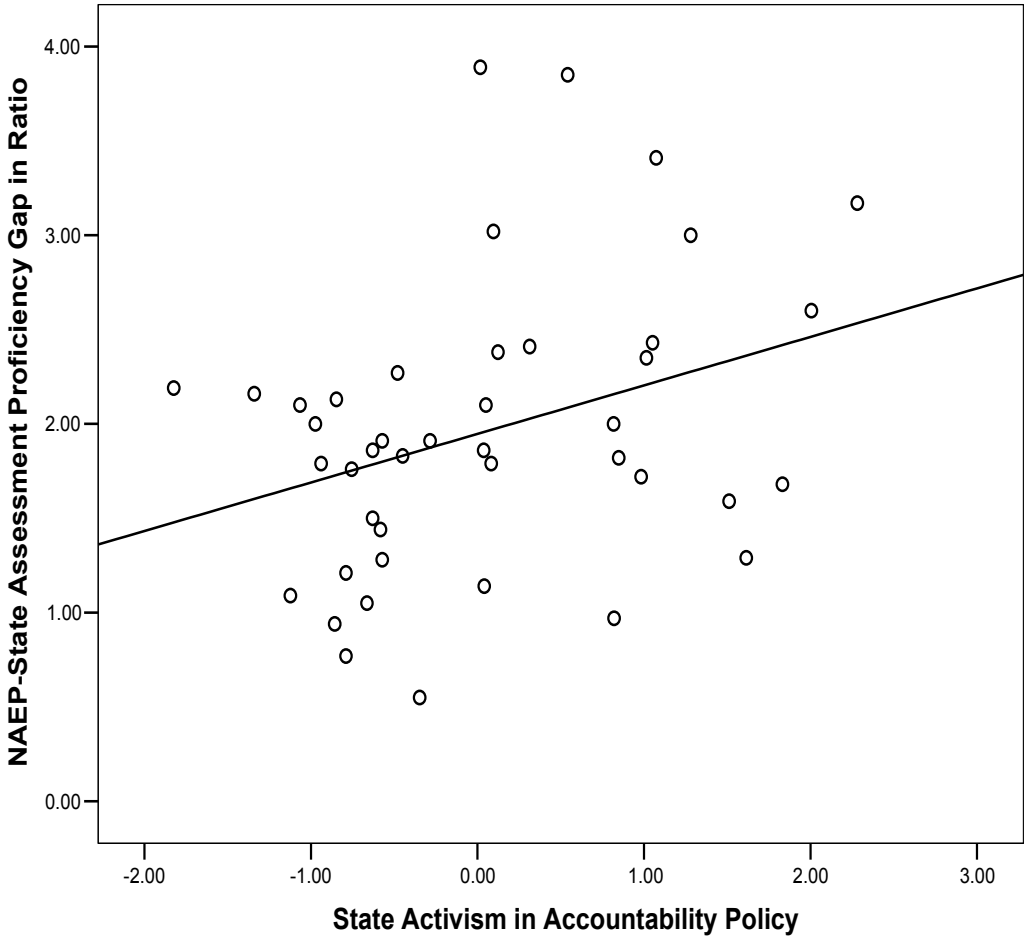

Figure 16: Plot of 50 States' Average Discrepancy between NAEP and State Assessment in the 8th Grade Math Proficiency (vertical axis) vs. Test-driven External Accountability Policy (horizontal axis)

Further, there are indications that states with high-stakes accountability systems show relatively smaller racial achievement gaps on their own state tests than on NAEP. The correlation analysis supports the hypothesis that high stakes testing tends to deflate

achievement gaps. The stronger test-driven external accountability, the smaller the discrepancies between NAEP and state assessment for achievement gaps, particularly White-Black gap in math. For example, the correlation between the level of state accountability and the size of discrepancy between NAEP and state assessment in grade 8 math White-Black gap was significantly negative (r = -.36).

NAEP vs. State Assessment Results on Post-NCLB Proficiency Gains

There is also the possibility that there is a discrepancy between the state assessment results and NAEP results in the amount of academic progress students made before NCLB and after NCLB. Unfortunately, currently available state assessment data are limited in their time span (typically available for up to the last 3-5 years) so that it is not possible to trace the pre-NCLB trend in most cases. Therefore, only the post-NCLB trend was compared in this study.

To determine if student progress on state assessments differed from student progress on NAEP, average gains in statewide proficiency rates from 2003 to 2005 was calculated for each state for both NAEP and state assessments. There were 25 states that had 2005 state reading and math assessment results available on their state education department web sites by the end of 2005.[16] Figures 17 and 18 illustrate the discrepancies in grade 8 proficiency gain estimates in reading and math respectively based on NAEP and state assessments. It shows that the gain was greater on the state assessments than on NAEP for those 25 states. In both grades 4 and 8 reading and in grade 8 math, there was no progress or a slight decline on NAEP, whereas there was some positive gain on the state assessments. In grade 4 math, both assessments showed progress, but the size of gain was smaller on NAEP. Table C-10 in Appendix C summarizes the results by each subject and grade.

[16] This number includes states that gave state assessments to the same grades as NAEP (grade 4 and grade 8) or adjacent grades: Alaska, Arizona, California, Colorado, Delaware, Georgia, Hawaii, Idaho, Indiana, Kansas, Louisiana, Maine, Maryland, Massachusetts, Michigan, Mississippi, Missouri, Nevada, New Jersey, North Carolina, Ohio, Pennsylvania, South Dakota, Virginia, and Washington.

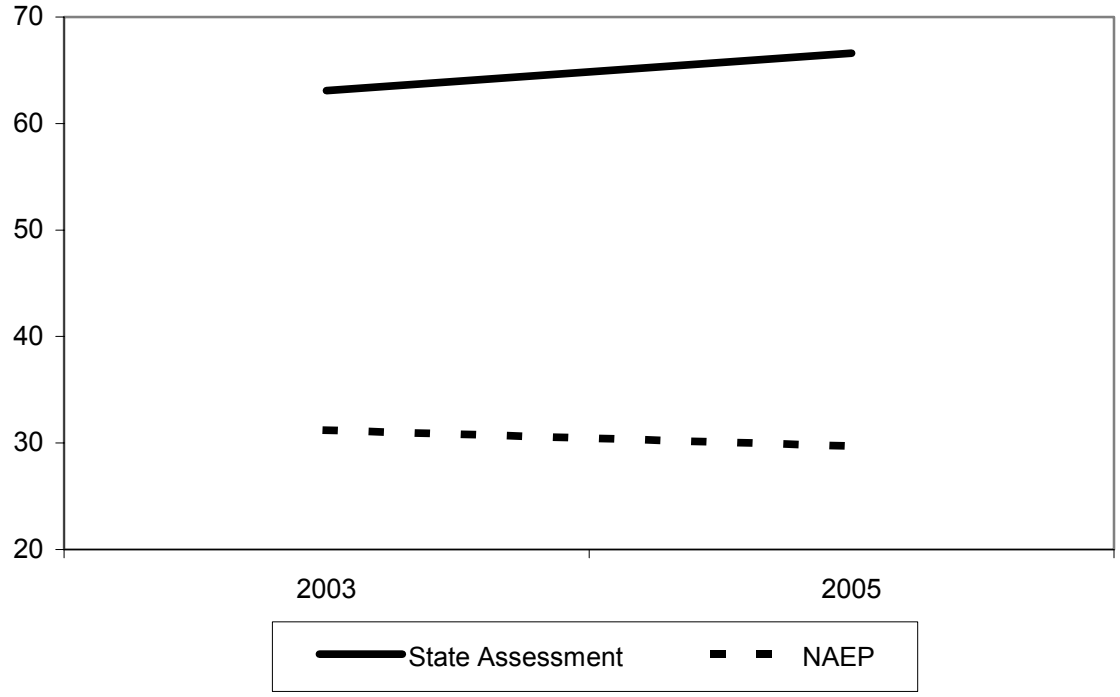

Figure 17: 2003-05 Grade 8 Reading Proficiency Trends based on State Assessment vs. NAEP (N = 25 states)

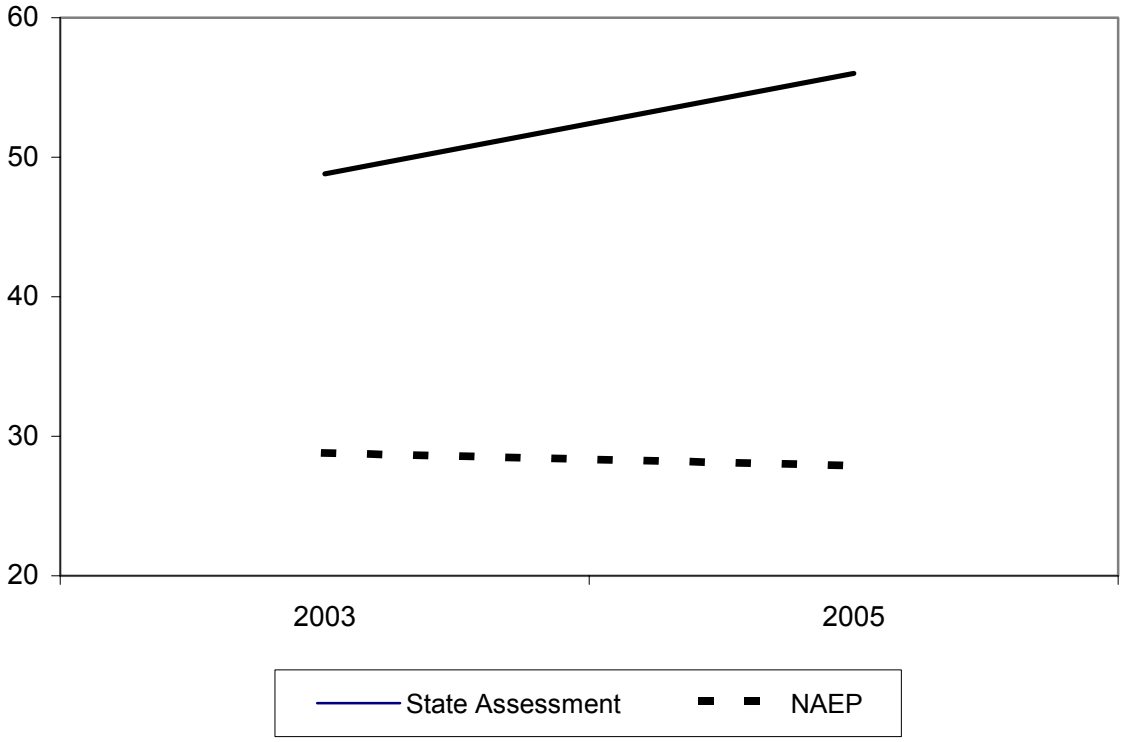

Figure 18: 2003-05 Grade 8 Math Proficiency Trend based on State Assessment vs. NAEP (N = 25 states)

Although there can be other reasons for these discrepancies between NAEP and state assessments in the size of estimated achievement gain scores, the gaps may be attributable partly to the fact that state assessment results are the basis of school accountability decisions under NCLB. High-stakes testing situations can lead to the possible inflation of achievement gains since schools may focus on teaching to the test as opposed to adopting changes that lead to genuine progress in learning. In the long term, NAEP and state assessment results may converge as a result of the increasing role of NAEP as a confirmatory tool and thus we may see greater alignment of state assessment results with NAEP under NCLB (Lee, in press-a).

PART 5: CONCLUSION

The goal of NCLB, which requires that states have all students accomplish high standards of learning in core subject areas (i.e., 100% of students become proficient in reading and math by 2014), is laudable. In the past, few states have been able to narrow racial and socioeconomic achievement gaps while improving overall achievement levels at the same time. If the law can facilitate the systemic efforts of state education systems to close pernicious achievement gaps, this would be noteworthy. Past and current NAEP reading and math achievement trends, however, raise serious concerns about the unrealistic performance goal and timeline and the possible consequences for schools that repeatedly fail to meet their performance target. If the nation continues to make the same amount of achievement gains as it did over the past 15 years, it may end up meeting only less than half of the reading proficiency target and less than two-thirds of the math proficiency target by 2014. These projections become much gloomier when it comes to closing the achievement gaps for disadvantaged minority students who are even more left behind in reading and math proficiency. However, it is worth noting that enormous progress was made in narrowing racial achievement gaps in the 1970s and 1980s (e.g., reduction of the Black-White math gap by half). This implies that further progress in closing the gap can be made through social and educational policies, reversing the setback in the 1990s.

In order to find out whether test-driven external accountability policy, the hallmark of NCLB, works, we need to know how well the nation and states have accomplished the goals of academic excellence and equity before NCLB as well as after NCLB. What do we learn from comparisons of pre-NCLB vs. post-NCLB NAEP trends of reading and math achievement? The results of national NAEP trend analyses suggest that NCLB did not have significant impact on improving reading and math achievement across the nation and states so far. The national average achievement remains flat in reading and grows at the same pace in math after NCLB that it did prior to NCLB. It is misleading to claim that NCLB has a positive effect on academic achievement simply because the national average math test scores continue to rise after NCLB. This inference is flawed because the increase in NAEP scores was just part of trend that began before NCLB and does not reflect any significant acceleration in the pace of academic improvement after NCLB. Nevertheless, it can also be misleading to discredit any potential effects of NCLB on achievement gains in the future by simply looking at the overall national growth trend in such a relatively short time period. Since some states had implemented their own school accountability systems long before NCLB, the impact of NCLB on individual states may be uneven or obscured by looking at the national aggregate picture.

By and large, the results of state-level NAEP trend analyses imply that NCLB's attempt to scale up the alleged success of the first generation accountability states (e.g., Florida, North Carolina, Texas) have so far not been effective. NCLB neither enhanced the first generation states' earlier academic improvement nor transferred the effects of their test-driven accountability policy to the second generation accountability states. The first generation accountability states made relatively greater academic progress before

NCLB in math but not in reading. Moreover, the relatively larger math gains among the first generation states were not sustained after NCLB. More importantly, states that adopted test-driven external accountability either before or after NCLB did not reduce racial and socioeconomic inequalities in reading and math achievement. It is evident that test-driven external accountability, whether it was a state or federal initiative, has not advanced equity on a large scale, as the disparity in achievement among different racial and socioeconomic groups of students persists before and after NCLB.

Another approach to closing the achievement gap may focus on each racial or socioeconomic subgroup's performance relative to a desired proficiency standard. In this view, the progress of Black, Hispanic, or Poor student subgroups towards the standard is evaluated on its own merits, and passing a designated threshold will be treated as narrowing the gap regardless of how the subgroup does in comparison to the White or Nonpoor subgroups. This approach is implied in NCLB, as the goal is to have every student meet a desired performance standard. If we adopt such a criterion-referenced view of closing the gap, we may observe a subgroup continuing to make progress toward the proficiency goal. However, meeting a pre-set threshold conveys no information about the achievement gap, and only provides information on whether or not a particular goal is reached. If the proficiency level is set so that the vast majority of Whites are over it at the beginning of the period studied, any improvement of minority students' proficiency rate may misleadingly signify progress toward closing the gap and obscure the relative gap between racial groups. This is of particular concern since NCLB establishes state assessments as the basis for school accountability and state standards vary widely in relationship to NAEP standards.

Despite the increasing importance of NAEP as the source of information for national and state report cards, the current practice of using states' own student assessments for school accountability purposes requires us to investigate the adequacy and utility of both assessments. In comparison with NAEP, state assessments tend to inflate the overall proficiency level and at the same time deflate the achievement gap among racial groups. This poses a threat to the validity of inferences based solely on states' own standards and assessment results. The results imply that the first generation accountability states with high-stakes testing policies in place prior to NCLB have adopted relatively lower performance standards, leading to overestimation of their proficiency rates and underestimation of the achievement gap. The findings also suggest that policy-makers become more aware of potential biases resulting from relying exclusively on states' own test measure for accountability.

It is time to reexamine the law, particularly in light of the evidence on the inefficacy of current test-driven external accountability policy to address the achievement gap under NCLB. While failure is not an option in education, it is important to acknowledge the limitations of the current policy and find solutions to problems that may have impeded national and state progress towards academic excellence and equity. It appears that NCLB follows the right path by combining input-guarantee and performance-guarantee approaches: it requires not only high performance standards and high-stakes testing for every student but also highly qualified teachers in every classroom

and more evidence-based funding for curricular and instructional reforms. In practice, however, NCLB has shortchanged many states with under-funded mandates and an over reliance on sanctions rather than a focus on capacity building. There remain substantial variations among states in the definitions and levels of student proficiency and teacher qualification and the adequacy and equity of school resources. In addition, substantial disparities in educational opportunities among racial and socioeconomic groups within states have not been adequately addressed. If NCLB revises the current course of test-driven accountability with shifts to more realistic goals and greater support for disadvantaged high-minority schools and puts forth a series of systemic reform efforts for continued improvement of educational opportunities on the equity front, it may be that NCLB will produce more positive results.

REFERENCES

Adams, J. E., & Kirst, M. W. (1999). New demands and concepts for educational accountability: Striving for results in an era of excellence. In J. Murphy & K. S. Louis (Eds.), *Handbook of research on educational administration* (pp. 463–490). San Francisco: Jossey-Bass.

Ad Hoc Committee on Confirming Test Results (2002, March 1). *Using the National Assessment of Educational Progress to confirm state test results: A report of the Ad Hoc Committee on Confirming Test Results.* Washington, DC: National Assessment Governing Board. Retrieved December 19, 2005 from http://nagb.org.

Amrein, A. L. & Berliner, D.C. (2002, March 28). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved June 14, 2003 from http://epaa.asu.edu/epaa/v10n18/.

Barton, P. E. (2002). *Raising achievement and reducing gaps: Reporting progress toward goals for academic achievement in mathematics.* Washington, DC: National Education Goals Panel.

Bracey, G. W. (2002). Standards and Achievement Gaps. *Phi Delta Kappan, 83*(8), 643.

Braun, H. (2004). Reconsidering the impact of high-stakes testing, *Education Policy Analysis Archives, 12*(1). Retrieved March 10, 2004 from http://epaa.asu.edu/epaa/v12n1/.

Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006). The Black-White achievement gap: Do state policies matter? *Education Policy Analysis Archives, 14*(8). Retrieved May 1, 2006 from http://epaa.asu.edu/epaa/v14n8/.

Carnoy, M., Loeb, S., & Smith, T. (2001). *Do higher scores in Texas make for better high school outcomes?* CPRE Research Report No. RR-047. Philadelphia, PA: Consortium for Policy Research in Education.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Center on Education Policy (2006). *From the Capital to the Classroom: Year 4 of the No Child Left Behind Act.* Retrieved on May 15, 2006 from http://www.cep-dc.org.

Education Trust (2004). *Measured Progress: Achievement Rises and Gaps Narrow, But Too Slowly.* Washington, DC: Education Trust.

Education Trust (2006). *Primary Progress, Secondary Challenge: A State-by-state Look at Student Achievement Patterns.* Washington, DC: Education Trust.

Elmore, F., & Fuhrman, S. (1995). Opportunity-to-learn standards and the state role in education. *Teachers College Record, 96,* 432–457.

FairTest (2005, October 19). Flatline NAEP scores show failure of test-driven school reform: "NO CHILD LEFT BEHIND" has not improved academic performance. Press release. Retrieved on May 15, 2006 from http://www.fairtest.org.

Fuhrman, S. H. (1999). *The new accountability.* (CPRE Policy Brief Series RB-27). Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas.* Washington, DC: National Education Goals Panel.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: Rand.

Goertz, M. E., & Duffy, M. E. (2001). *Assessment and accountability systems in the 50 states: 1999–2000* (CPRE Research Report RR-046). Philadelphia, PA: Consortium for Policy Research in Education.

Haney. W. (2000). The myth of the Texas miracle in education. *Educational Policy Analysis Archives*. Retrieved March 3, 2001 from http://epaa.asu.edu/epaa/v8n41.

Hanushek, E. A., & Raymond, M. E. (2004). Does school accountability lead to improved performance? *Journal of Policy Analysis and Management*. 24(2), 297-327.

Harris, D. N. & Herrington, C. D. (2006). Accountability, Standards, and the Growing Achievement Gap: Lessons from the Past Half-Century. *American Journal of Education, 112,* 209–238

Henderson-Montero, D., Julian, M. W., & Yen, W. M. (2003). Multiple measures: Alternative design and analysis models. *Educational Measurement: Issues and Practice, 22*(2), 7-12.

Jencks, C., & Phillips, M. (Eds.). (1998). *The black-white test score gap*. Washington, D.C.: Brookings Institution Press.

Kim, J. S., & Sunderman, G. L. (2005). Measuring Academic Proficiency Under the No Child Left Behind Act: Implications for Educational Equity. *Educational Researcher*.

Klein, S. P., Hamilton, L.S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: Rand.

Koretz, D., & Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS)* (MR-792-PCT/FF). Santa Monica, CA: Rand.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31(1)*, 3–12.

Lee, J. (2003). Evaluating Rural Progress in Mathematics Achievement: Threats to the Validity of "Adequate Yearly Progress." *Journal of Research in Rural Education, 18*, 67-77.

Lee, J. (2004). How Feasible is Adequate Yearly Progress (AYP)? Simulations of School AYP "Uniform Averaging" and "Safe Harbor" under the No Child Left Behind Act. *Educational Policy Analysis Archives, 12*(14). Retrieved May 1, 2006 from http://epaa.asu.edu/epaa/v12n14/.

Lee, J. (2006). Is Test-driven External Accountability Effective? A Meta-analysis of the Evidence from Cross-State Causal-Comparative and Correlational Studies. Paper presented at the annual meeting of American Educational Research Association (AERA), San Francisco, California.

Lee, J (in press-a). Do National and State Assessments Converge for Educational Accountability? A Meta-Analytic Synthesis of Multiple Measures in Maine and Kentucky. *Applied Measurement in Education*.

Lee, J. (in press-b). Input-guarantee vs. Performance-guarantee Approaches to School Accountability: Cross-State Comparisons of Policies, Resources, and Outcomes. *Peabody Journal of Education*.

Lee, J., & Wong, K. K. (2004). The Impact of Accountability on Racial and Socioeconomic Equity: Considering both School Resources and Achievement Outcomes. *American Educational Research Journal, 41*, 797-832.

LeFloch, K. C., Taylor, J., & Thomsen, K. (2006). The implications of NCLB accountability for comprehensive school reform. Paper presented at the annual meeting of American Educational Research Association.
Linn, R. L. (2000). Assessments and accountability. *Educational Researcher. 2*(29), 4-16.
Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3–13.
Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31(6)*, 3-16.
Linton, T. H. & Kester, D. (2003). Exploring the achievement gap between white and minority students in Texas: A comparison of the 1996 and 2000 NAEP and TAAS eighth grade mathematics test results, Education Policy Analysis Archives, 11(10). Retrieved May 1, 2006 from http://epaa.asu.edu/epaa/v11n10/.
Mathis, W. J. (2003). No Child Left Behind: Costs and Benefits. *Phi Delta Kappan, 84*(9), 679-686. Retrieved 16 March, 2003 from http://www.pdkintl.org/kappan/k0305mat.htm
Mintrop, H. & Trujillo, T.M. (2005). Corrective action in low performing schools: Lessons for NCLB implementation from first-generation accountability systems. *Education Policy Analysis Archives, 13*(48). Retrieved December 15, 2005 from http://epaa.asu.edu/epaa/v13n48/.
Mullis, I. V.S. et al. (1993). *NAEP 1992 Mathematics Report Card for the Nation and the States.* Washington, D.C.: National Center for Education Statistics. Report No. 23-ST02.
National Education Goals Panel (1996). *Profile of 1994-95 state assessment systems and reported results.* Washington, DC: Author.
NAACP (2005). *Moving From Rhetoric To Reality In Opening Doors To Higher Education for African-American Students.* New York: Author.
National School Boards Association (2006). *Federal funding for education.* Alexandria, VA: Author.
Nichols, S. L., Glass, G. V, & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved May 1, 2006 from http://epaa.asu.edu/epaa/v14n1/.
No Child Left Behind Act of 2001, Pub. L. No. 107-110.
Newmann, F. M., King, M. B., & Rigdon, M. (1997). Accountability and school performance: Implications from restructuring schools. *Harvard Educational Review, 67*(1), 41–74.
O' Day, J. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review, 72*(3).
O'Day, J., & Smith, M. (1993). Systemic reform and educational opportunity. In S. Fuhrman (Ed.), *Designing coherent educational policy* (pp. 250–312). San Francisco: Jossey-Bass.
Perie, M., Grigg, W., & Dion, G. (2005). *The nation's report card: Mathematics 2005* (NCES 2006-453). U.S. Department of Education, NCES. Washington, D.C.: U.S. Government Printing Office.

Perie, M., Grigg, W., & Donahue, P. (2005). *The nation's report card: Reading 2005* (NCES 2006-451). U.S. Department of Education, NCES. Washington, D.C.: U.S. Government Printing Office.

Peterson, P. E. (Ed.) (2006). *Generational change: Closing the test score gap.* Lanham, MD: Rowman & Littlefield.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis* (2nd ed.). Newbury Park, CA: SAGE.

Raymond, M. E., & Haushek, E. A. (2003). High-stakes research. *Education Next,* Summer/No.3. Retrieved January 20, 2004 from http://www.educationnext.org/20033/index.html.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the Black-White achievement gap.* Washington, DC: Economic Policy Institute.

Singer, J.D. & J.B. Willett. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Skrla, L., Scheurich, J. J., Johnson, J. F., & Koschoreck, J. W. (2004). Accountability for equity: Can state policy leverage social justice? (Ch. 5, pp. 51-78) In Skrla, L. & Scheurich, J. J. (Eds.) *Educational Equity and Accountability: Paradigms, policies, and politics.* New York: RoutledgeFalmer

Sunderman, G. L., Kim, J. S., & Orfield, G. (2005). *NCLB meets school realities: Lessons from the field*. Thousand Oaks, CA: Corwin Press.

U.S. Department of Education (2005), *The Achiever, 4*(12).

Valencia, R. R., Valenzuela, A., Sloan, K., & Foley, D. (2004). Let's treat the cause, not the symptoms: Equity and accountability in Texas revisited. (Ch. 3, pp. 29-38) In Skrla, L. & Scheurich, J. J. (Eds.) *Educational Equity and Accountability: Paradigms, policies, and politics.* New York: RoutledgeFalmer

West, M. R., & Peterson, P. E. (2005). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. Paper presented at the Annual Conference of the Royal Economic Society, University of Nottingham.

APPENDIX A. DATA AND STATISTICAL METHODS

Data

NAEP provides repeated cross-sectional measures of reading and math achievement for each grade. The NAEP results are reported in two ways: scale scores and the percentages of students scoring at or above three benchmarks called achievement levels (Perie, Grigg, & Donahue, 2005 for reading; Perie, Grigg, & Dion, 2005 for math). NAEP reading and math scores are on a 0-500 scale. Interpretation of the NAEP scale scores is made with reference to performance standards for each subject and grade, using corresponding cut scores for three achievement levels: Basic, Proficient and Advanced.

This study used national-level and state-level aggregate measures of performance in scale scores and the percentages of students scoring at or above Proficient level that were drawn from 1990-2005 NAEP public school sample grade 4 and grade 8 reading and math assessments (www.nces.ed.gov/nationsreportcard). The NAEP national grade 4 and 8 data were drawn from the NAEP database for the following years: 1992, 1994, 1998, 2000 (grade 4 only), 2002, 2003, 2005 in reading and 1990, 1992, 1996, 2000, 2003, 2005 in math. The NAEP state grade 4 and 8 data were drawn from the NAEP database for the following years: 1992 (grade 4 only), 1994 (grade 4 only), 1998, 2002, 2003, 2005 in reading and 1990 (grade 8 only), 1992, 1996, 2000, 2003 and 2005 in math. Grade 12 was not included in the study due to the lack of available data. The NAEP 2005 national results for grade 12 were not available at the time of this study so that it was not possible to assess post-NCLB trend at the high school level. There are also no NAEP state assessment data at grade 12 or any other high school grades.

Since 1998 in reading and 1996 in math, testing accommodations (e.g., extended testing time, individual test administration) were provided to students with disabilities and/or English language learners. Therefore, the NAEP results with accommodation permitted were used for the 1998-2005 years in reading and for the 1996-2005 years in math. All prior assessment results were without accommodation. For the sake of keeping track of achievement throughout the 1990s prior to NCLB, all available NAEP data points, including results with and without accommodation, were used. Preliminary analysis for this study attempted to adjust the national or state achievement trends for changes in the accommodation policy but did not detect significant bias in the estimation of pre-NCLB achievement trends.

To analyze the racial achievement gap on NAEP, the average achievement of Black and Hispanic students was compared with the average achievement of White students. Although the NAEP data analysis includes Asian and Pacific Islanders as well, the analysis of racial gaps focused on the achievement of Blacks and Hispanics who have significant gaps relative to their White counterparts. To analyze the socioeconomic achievement gap, comparisons were made between Poor and Nonpoor students as classified by eligibility for free or reduced-price lunch. The NAEP data broken down by this school lunch variable for Poor and Nonpoor students and their achievement gap are not available until 1998 in reading and 1996 in math.

Interpretation of the achievement gap on NAEP can be facilitated by using some sort of effect size metrics. One way to think about the size of the achievement gap is considering how large the gap is relative to the standard deviation of NAEP scores. In order to compute standardized gap scores at grades 4 or grade 8, the gap score can be divided by within-grade standard deviation of student scores. The distributions of student scores at the baseline year across the national public school sample are as follows: M=215, SD=36 for 1992 grade 4 reading; M=258 SD=36 for 1992 grade 8 reading; M=212, SD=32 for 1990 grade 4 math; M=262, SD=36 for 1990 grade 8 math. Another way to think about the size of the achievement gap is to consider how large the gap is relative to the average amount of gain score per grade on the NAEP scale (about 10-12 point gain per grade based on the difference between grade 4 and grade 8 average scores).

Statistical Methods

Weighted Least Squares (WLS) regression was used to analyze the national trends of reading and math scores in PART 2 and takes into account the precision of national average scores, gaps, or proficiency rates. Weight was calculated by taking the inverse of standard errors of average scores or proficiency estimates for each group and the gap between groups at each time point. More recent assessments tend to have smaller standard errors. The entire period for which NAEP data is available was divided into two periods, Pre-NCLB (1990-2001) and Post-NCLB (2002-2005). The following two-piece linear growth model postulates a national academic growth trajectory with two temporal predictors of outcome Y. It affords testing whether there was significant increment or decrement to the baseline growth rate after NCLB:

$$Y_t = \pi_0 + \pi_1(\text{Pre-NCLB})_t + \pi_2(\text{Post-NCLB})_t + e_t$$

Where

Yt is the measure of nation's average achievement outcome at year t;

(Pre-NCLB)t is the number of years elapsed since the first NAEP assessment at year t (0 for 1990, 1 for 1991, ….., 15 for 2005);

(Post-NCLB)t is the number of years elapsed since the enactment of NCLB at year t (0 for 1990 through 2001, 1 for 2002, 2 for 2003, ….., 4 for 2005);

π_0 is the initial status of achievement;

π_1 is pre-NCLB annual growth rate during the baseline time period (achievement gain per year during 1990-2001);

π_2 is post-NCLB increment or decrement to the baseline pre-NCLB growth rate (change in π_1 during 2002-05);

e_t is a random effect representing the deviation of nation's score from the predicted score based on the model.

Hierarchical linear models (HLM), two-piece linear growth models, were used in PART 3 to examine interstate variations in the trends of reading and math achievement over the 1990-2005 period (Raudenbush & Bryk, 2002). Since there were four outcome variables for each group (grade 4 reading, grade 4 math, grade 8 reading, and grade 8 math), 2-level HLM analyses were conducted separately for each outcome variable, using the precision of the outcome variable as weight. At Level 1 (time level), the same two

temporal predictors were used to keep track of each state i's outcome variable Y at year t. The level-1 coefficients, including initial status (π_{1i}), pre-NCLB growth rate (π_{1i}) and post-NCLB change in the growth rate (π_{2i}), were assumed to vary randomly among states. Also, the study's assumption of independent errors with constant variance is unlikely to distort the analysis for a short time series. At Level 2 (state level), state activism in test-driven external accountability policy was used as one of the predictors to account for these interstate variations in academic growth patterns (see Appendix B for description of the Accountability variable). Further, HLM latent variable regression method was used to control for the effect of initial status on pre-NCLB gain as well as the effect of both initial status and pre-NCLB growth rate on post-NCLB change.

Level 1 Model:

$$Y_{ti} = \pi_{0i} + \pi_{1i}(\text{Pre-NCLB})_{ti} + \pi_{2i}(\text{Post-NCLB})_{ti} + e_{ti}$$

Level 2 Model:

$$\pi_{0i} = \beta_{00} + \beta_{01}(\text{Accountability})_i + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11}(\text{Accountability})_i + \beta_{12}(\pi_{0i}) + r_{1i}$$

$$\pi_{2i} = \beta_{20} + \beta_{21}(\text{Accountability})_i + \beta_{22}(\pi_{0i}) + \beta_{23}(\pi_{1i}) + r_{2i}$$

It needs to be noted that the above growth model is simply one of several possible models since there are other alternative growth models (Singer & Willet, 2003). The above model postulates a discontinuity in slope, not elevation; it is hypothesized that the growth rate changes after NCLB. In contrast, a model can include a discontinuity in elevation, not slope; it means that a temporary change right after NCLB is followed by a return to the pre-reform growth rate. This alternative model was also tested separately and the results from grade 4 reading and math only provided limited support for the model with the NAEP average grade 4 reading scores for All, White, Nonpoor and grade 4 math scores for Black and Hispanic. Testing a model with a discontinuity in both elevation and slope together was not possible due to insufficient post-NCLB data points.

One advantage of the multilevel model for change is that it improves the precision of estimation of individual growth parameters. These model-based estimates of growth trajectories combine Ordinary Least Squares (OLS) estimates with population average estimates derived from the fitted models. This combination yields a superior, more precise, estimate when data are sparse. For instance, in this study, there may be too few data points in some states to enable valid statistical inferences on the average proficiency or gap trend using traditional regression models. HLM models can use not only the data in those short-term states but also information in the pooled data for all states, including long-term ones. Therefore, the pooling involved in multilevel models affords a "borrowing of strength" that supports statistical inference in a situation where no inference would be possible using traditional methods (Raudenbush & Bryk, 2002; Singer & Willet, 2003). This HLM analysis provided for testing statistical significance of the growth rate in each state. Statistical significance of each state's pre-NCLB growth and post-NCLB change was determined by using a more rigorous alpha level of .001,

which controls for familywise Type I errors in testing the same set of hypotheses with all 50 states.

It needs to be noted that this study involved tracking successive cohort groups of students at the same grade over time. This grade-based (repeated cross-sectional) comparison method that tracks test score changes for the same grade (e.g., 1996 8^{th} grade to 2000 8^{th} grade) contrasts with a cohort-based (quasi-longitudinal) comparison method which tracks the performance of the same cohort group (e.g., 1996 4^{th} grade to 2000 8^{th} grade). Review of previous studies on the impact of accountability on achievement revealed contradictory results between the two methods; the grade-based comparison method tended to produce more positive results whereas the cohort-based comparison method showed more negative effects (Lee, 2006). Examination of the post-NCLB achievement trend through the cohort-based method is not possible yet, since the currently available post-NCLB NAEP data does not afford a 4-year interval between the measures.

Another potential factor that may confound the results of the average achievement and gap trend analysis is change in the identification and exclusion of certain groups of students for NAEP testing, particularly students with learning disabilities (SWD) and English language learners (ELL). Increasing number of ELL students particularly among Hispanic and Asian immigrant populations, could have influenced the average Hispanic and Asian achievement trends. On one hand, as a result of demographic changes, the national average identification rate of SWD and/or ELL students in NAEP has increased over the past 15 years and thus tends to be higher for the post-NCLB period than for the pre-NCLB period. On the other hand, as a result of accommodation permitted since 1996 in math and since 1998 in reading, the national average exclusion rate of SWD and/or ELL students in NAEP has decreased over time and thus tends to be somewhat lower for post-NCLB period than for pre-NCLB period. Preliminary analysis of this study showed that these factors do not significantly affect findings on the national trends of reading and math achievement during the post-NCLB period.

Since the exclusion rate of SWD and/or ELL students varied from state to state, we also need to consider this interstate variation for a fair comparison of the state achievement trends. Amrein and Berliner (2002) point out that the larger achievement gains in high-stakes testing states such as North Carolina and Texas are attributable partly to their relatively large increases in exclusion rates. However, Braun (2004) showed that those two states are outliers that deviate from the pattern of weak or no relationship between change in exclusion rate and gain scores among all participating NAEP states. Carnoy and Loeb (2002), Raymond and Haushek (2003), and Hanushek and Raymond (2004) studies also show that statistically adjusting gain scores for changes in exclusion rates did not lead to significant changes in the estimation of accountability policy effects. Preliminary analysis of this study also did not find significant influence of exclusion rates on state accountability policy effect estimates.

APPENDIX B. MEASURES OF STATE ACCOUNTABILITY AND THE DISCREPANCIES BETWEEN NAEP AND STATE ASSESSMENT IN READING AND MATH PROFICIENCY

State Activism in Test-driven External Accountability

This report utilizes the measures of test-driven external accountability policy for 50 states as constructed by Lee and Wong (2004). It is based on survey data collected in the mid to late 1990s from three sources: (1) 1995-96 data from the North Central Regional Education Laboratory (NCREL) and Council of Chief State School Officers (CCSSO) (NCREL/CCSSO, 1996), (2) 1999 data from Quality Counts (QC) report (Education Week, 1999), and (3) 1999-2000 data from the Consortium for Policy Research in Education (CPRE) report (Goertz & Duffy, 2002). The NCREL/CCSSO survey covers student assessments, student accountability (testing for promotion, awards/recognition, and graduation), teacher accountability (certification gain/loss, financial rewards/penalties, probation), and school accountability (funding gain/loss, accreditation loss, awards/recognition, performance reporting, probation/warning, takeover/dissolution). The QC survey covers only student assessments and school accountability (report cards, ratings, rewards, assistance and sanctions). The CPRE survey covers student assessments and student and school accountability policies (school/district sanctions or rewards, high school exit test).

Here are some sample questions and response options from the NCREL/CCSSO survey:

1. What uses are made of the results of the assessment for student accountability? (1) Student awards or recognition, (2) Promotion, (3) Honors diploma, (4) Endorsed diploma, (5) Graduation.

2. What uses are made of the results of the assessment for school accountability? (1) School awards or recognition, (2) School performance reporting, (3) High school skills guarantee, (4) School accreditation.

3. Does this assessment have consequences for schools? (1) Funding gain, (2) Exemption from regulations, (3) Warnings, (4) Probation, watch lists, (5) Funding loss, (6) Accreditation loss, (7) Takeover, (8) Dissolution.

4. Does this assessment have consequences for school staff? (1) Financial rewards, (2) Certification status gain, (3) Probation, (4) Certification status loss, (5) Financial penalties.

Policy index scores were calculated for each state by summing the number of policies adopted and in place by the state at the time of survey. The NCREL/CCSSO policy index ranges from zero to 16 (M = 6.5, SD = 4.2). The reliability of this 26-item 1995 NCREL/CCSSO accountability policy index is very high (alpha = .85). The QC policy index ranges from zero to 6 (M = 3.0, SD = 1.8). The reliability of this 6-item 1999 QC accountability policy index is high (alpha = .77). Finally, the CPRE policy index was constructed by Carnoy and Loeb (2002) and it ranges from zero to five (M = 2.1, SD = 1.4).

Out of these three related policy measures, Lee and Wong (2004) created a composite factor of state activism in test-driven external accountability policy during the 1990s. One factor is retained through principal component analysis of the three state-level policy index variables with high factor loadings: '95 NCREL/CCSSO policy index, .85; '99 Quality Counts policy index, .87; and 2000 CPRE policy index, .85. Factor has an eigen value of 2.2 and explains 74 percent of the combined variance. Table B-1 shows the measures of state activism in accountability for all states.

Discrepancies between NAEP and State Assessment in Reading and Math Proficiency

In PART 4, the NAEP assessment results for individual states were compared with states' own assessment results in 4th and 8th grade reading and math. Since state assessment results were most readily available in the form of the percentage of students who meet a desired standard (typically at or above a Proficient level), proficiency rate data were obtained from each of the 43 state education departments that made this data available on their websites and were matched to corresponding NAEP proficiency rates in the same subject and grade during the same testing year. When all the available data are stacked across multiple years and states, the numbers of maximum data points are as follows: N=90 in grade 4 math, N=115 in grade 4 reading, N=103 in grade 8 reading, N=82 in grade 8 math. The number of states varies by year: in grade 4 math for example, N=3 in 1996, N=17 in 2000, N=45 in 2003, N=25 in 2005.

Table B-1 shows the measures of NAEP vs. state assessment discrepancies in each grade and subject across years for 43 states that have both NAEP and state assessment results available. The discrepancy between the two assessments was measured by the ratio of the state assessment-based proficiency rate to the NAEP-based proficiency rate. The more this ratio departs from the value of one, the greater the discrepancies between the two assessments. A ratio exceeding 1 implies a relatively lower state standard in comparison with the NAEP standard, whereas a ratio falling below 1 implies a relatively higher state standard.

Table B-1.
Measures of State Accountability and NAEP vs. State Assessment Discrepancies in Reading and Math Proficiency

State	Test-driven External Accountability Score	Ratio of State Assessment to NAEP Proficiency			
		Grade 4 Reading	Grade 4 Math	Grade 8 Reading	Grade 8 Math
Alabama	1.34	.	.	.	.
Alaska	-0.85	2.78	2.31	2.8	2.13
Arizona	-0.58	2.71	2.09	2.52	1.44

Arkansas	-0.79	.	.	.	.
California	0.04	1.94	1.79	1.61	1.14
Colorado	-1.34	2.38	2.42	2.54	2.16
Connecticut	-0.48	1.6	2.3	2.08	2.27
Delaware	-0.94	2.44	2.21	2.43	1.79
Florida	1.05	1.96	1.74	1.68	2.43
Georgia	0.1	3.04	2.96	3.17	3.02
Hawaii	-0.66	2.25	0.9	1.92	1.05
Idaho	-1.07	2.56	2.37	2.43	2.1
Illinois	0.85	1.94	2.44	1.83	1.82
Indiana	1.01	2.29	2	2.13	2.35
Iowa	-1.82	2.19	2.14	1.93	2.19
Kansas	0.08	2.12	1.92	2	1.79
Kentucky	1.61	2.02	1.73	1.71	1.29
Louisiana	1.07	3	2.93	2.35	3.41
Maine	-0.79	1.43	0.94	1.17	0.77
Maryland	1.83	1.62	1.95	1.43	1.68
Massachusetts	-0.86	1.23	0.9	1.56	0.94
Michigan	0.82	2.45	1.92	2.26	2
Minnesota	-0.49	1.88	1.83	.	.
Mississippi	0.02	5.01	4.26	2.76	3.89
Missouri	-0.35	1.05	1.41	0.99	0.55
Montana	-0.97	2.16	2.42	1.93	2
Nebraska	-1.82	.	.	.	.
Nevada	0.12	2.2	2.06	2.31	2.38
New Hampshire	-1.07	1.91	1.84	.	.
New Jersey	0.98	2.06	1.65	1.94	1.72

New Mexico	1.28	2.37	3	2.55	3
New York	1.51	1.81	2.73	1.33	1.59
North Carolina	2.01	2.58	2.63	2.96	2.6
North Dakota	-0.79	2.24	1.71	1.86	1.21
Ohio	0.05	2.06	1.58	2.45	2.1
Oklahoma	0.54	2.83	3.39	2.74	3.85
Oregon	-0.45	2.74	2.7	1.69	1.83
Pennsylvania	-0.63	1.72	1.62	1.81	1.86
Rhode Island	-0.63	2.05	1.55	1.43	1.5
South Carolina	0.82	1.28	1.2	1	0.97
South Dakota	-0.76	2.61	2.07	2.13	1.76
Tennessee	0.2	.	.	.	.
Texas	2.28	3.11	3.03	3.14	3.17
Utah	-0.57	2.43	2.33	2.12	1.91
Vermont	-0.29	2.13	1.74	1.59	1.91
Virginia	0.31	2.25	2.25	1.95	2.41
Washington	-0.57	1.99	1.49	1.47	1.28
West Virginia	0.71	.	.	.	.
Wisconsin	0.04	2.45	2.09	2.14	1.86
Wyoming	-1.12	1.36	0.95	1.2	1.09

APPENDIX C. SUPPORTING TABLES

Table C-1.
National Trends in NAEP Grade 4 Reading and Math Achievement by Subgroups and their Gaps

	Reading		Math	
	Pre-NCLB Growth	Post-NCLB Change	Pre-NCLB Growth	Post-NCLB Change
All	.13	.57	1.31*	.93
White	.46	.07	1.50*	.60
Black	.84	.36	1.83*	1.39
Hispanic	.49	1.33	1.29	1.98
Asian	.82**	.29	.91	2.67
Nonpoor	1.56	1.57	1.43	.87
Poor	1.96	-1.13	1.56	1.57
White-Black gap	-.44	-.18	-.39	-.62
White-Hispanic gap	-.12	-1.06	.17	-1.31
Poverty gap	-.96	.52	-.16	-.52

Note. Pre-NCLB growth column shows the estimate of national average yearly growth rate in each variable during 1990-2001 period, that is, gain or loss in the average score or gap per year. Post-NCLB change column shows the estimate of national average yearly change during the 2002-2005 period, that is, increment or decrement to the pre-NCLB growth rate as reported in its previous column for the same variable. Asterisks indicate statistical significance level of the estimate: * $p < .05$; ** $p < .01$

Table C-2.
National Trends in NAEP Grade 8 Reading and Math Achievement by Subgroups and their Gaps

	Reading		Math	
	Pre-NCLB Growth	Post-NCLB Change	Pre-NCLB Growth	Post-NCLB Change
All	.67*	-1.43*	.90**	.25
White	.82**	-1.27*	1.25**	-.61
Black	1.20**	-1.99*	1.03*	1.22
Hispanic	.70*	-.61	.74**	1.26*
Asian	-.08	1.81	-.53	2.61
Nonpoor	1.19*	-1.72*	1.69	-.88
Poor	1.17	-1.74	.95	.81
White-Black gap	-.40	.75	.18	-1.74
White-Hispanic gap	.11	-.66	.40	-1.70*
Poverty gap	.07	-.07	.75	-1.71

Note. Pre-NCLB growth column shows the estimate of national average yearly growth rate in each variable during 1990-2001 period, that is, gain or loss in the average score or gap per year. Post-NCLB change column shows the estimate of national average yearly change during the 2002-2005 period, that is, increment or decrement to the pre-NCLB growth rate as reported in its previous column for the same variable. Asterisks indicate statistical significance level of the estimate: * $p < .05$; ** $p < .01$

Table C-3.
National Trends in NAEP Grade 4 Reading and Math Proficiency by Subgroups and their Gaps

	Reading		Math	
	Pre-NCLB Growth	Post-NCLB Change	Pre-NCLB Growth	Post-NCLB Change
All	.009	.008	.069*	.059
White	.020*	.000	.083*	.068
Black	.037	.011	.144*	.084
Hispanic	.030**	.017	.068*	.178
Asian	.062	-.064	.019	.302
Nonpoor	.016	.007	.101	.059
Poor	.095	-.092	-.002	.281
White-Black gap	-.017	-.012	-.062	-.016
White-Hispanic gap	-.009**	-.017*	.015	-.110
Poverty gap	-.079	.099	.103	-.222

Note. Numbers are in a logit metric, that is, log odds of percent students performing at or above Proficient level. Pre-NCLB growth column shows the estimate of national average yearly growth in each variable during 1990-2001 period, that is, gain or loss in proficiency rate or gap per year. Post-NCLB change column shows the estimate of national average yearly change during the 2002-2005 period, that is, increment or decrement to the pre-NCLB growth rate as reported in its previous column for the same variable. Asterisks indicate statistical significance level of the estimate: * $p < .05$; ** $p < .01$

Table C-4.
National Trends in NAEP Grade 8 Reading and Math Proficiency by Subgroups and their Gaps

	Reading		Math	
	Pre-NCLB Growth	Post-NCLB Change	Pre-NCLB Growth	Post-NCLB Change
All	.027**	-.057*	.057*	-.048
White	.035**	-.060*	.073*	-.061
Black	.063***	-.117**	.039	.105
Hispanic	.029**	-.031*	.026	.098
Asian	-.014	.118	-.017	.083
Nonpoor	.049	-.076	.057*	-.018
Poor	.077	-.119	.053	.009
White-Black gap	-.028*	.057	.033	-.166
White-Hispanic gap	.006	-.029	.046	-.160
Poverty gap	-.027	.043	.004	-.027

Note. Numbers are in a logit metric, that is, log odds of percent students performing at or above Proficient level. Pre-NCLB growth column shows the estimate of national average yearly growth in each variable during 1990-2001 period, that is, gain or loss in proficiency rate or gap per year. Post-NCLB change column shows the estimate of national average yearly change during the 2002-2005 period, that is, increment or decrement to the pre-NCLB growth rate as reported in its previous column for the same variable. Asterisks indicate statistical significance level of the estimate: * $p < .05$; ** $p < .01$

Table C-5.
State Trends in NAEP Grade 4 Reading and Math Achievement by Subgroups and their Gaps (N=50 states)

	Reading		Math	
	Pre-NCLB Growth	Post-NCLB Change	Pre-NCLB Growth	Post-NCLB Change
All	M=.27 SD=.33 5 ↑ 45 —	M= .18 SD=.20 50 —	M= .96 SD= .35 39 ↑ 11 —	M= 1.70 SD= .14 50 ↑
White	M=.39 SD=.35 16 ↑ 34 —	M= .07 SD=.26 1 ↑ 49—	M=1.03 SD=.28 49 ↑ 1 —	M= 1.57 SD=.10 50↑
Black	M=.58 SD=.26 2 ↑ 39 —	M= .31 SD=.24 41 —	M=1.57 SD=.40 38 ↑ 3 —	M=1.44 SD=.57 20 ↑ 21 —
Hispanic	M=.91 SD=.82 2 ↑ 39 —	M= -.05 SD=.99 41 —	M=1.33 SD=.51 15 ↑ 27 —	M=1.93 SD=.52 25 ↑ 17—
Asian	M=.87 SD=.46 27 —	M=.60 SD=.44 27 —	M= 1.43 SD=.85 1 ↑ 26—	M=2.19 SD=1.35 27 —
Nonpoor	M=1.15 SD=.99 9 ↑ 41 —	M=-1.06 SD=.94 42 — 8 ↓	M=1.18 SD=.48 23 ↑ 27 —	M=1.24 SD=.30 43 ↑ 7 —
Poor	M=2.08 SD=1.24 10 ↑ 40 —	M=-2.11 SD=1.06 36 — 14 ↓	M=1.49 SD=.57 23 ↑ 27 —	M= 1.40 SD=.40 25 ↑ 25 —
White-Black gap	M=-.07 SD=.30 41 —	M=-.33 SD=1.05 41 —	M=-.37 SD=.36 40 — 1 ↓	M=-.18 SD=.73 41 —
White-Hispanic gap	M=-.22 SD=.46 40 — 1 ↓	M=-.20 SD=.22 41 —	M=-.12 SD=.24 42 —	M=-.44 SD=.35 40 — 2↓
Poverty gap	M=-1.00 SD=.45 50 —	M=1.18 SD=.65 50 —	M=-.24 SD=.41 50 —	M=-.27 SD=.53 50 —

Note. M is the mean of all states' growth parameter estimates, and SD is the standard deviation of the estimates across states. Statistical significance of each individual state's growth was determined at the .001 level to reduce familywise Type I error for simultaneous comparisons of multiple states. Numbers in front of arrows or dashes indicate the number of states for each pattern (↑ significantly up; ↓ significantly down; — no significant change).

Table C-6.
State Trends in NAEP Grade 8 Reading and Math Achievement by Subgroups and their Gaps (N=50 states)

	Reading		Math	
	Pre-NCLB Growth	Post-NCLB Change	Pre-NCLB Growth	Post-NCLB Change
All	M=.65 SD=1.15 2 ↑ 48 —	M= -1.16 SD=1.40 49 — 1 ↓	M=.90 SD=.40 34 ↑ 16 —	M=-.14 SD=.81 50 —
White	M=.80 SD=.88 3 ↑ 47 —	M= -1.26 SD=1.08 49 — 1 ↓	M= 1.02 SD=.41 38 ↑ 12 —	M= -.14 SD=.82 50 —
Black	M=.71 SD=1.54 1 ↑ 40 —	M=-1.08 SD=1.91 41 —	M=1.06 SD=.47 15 ↑ 25 —	M=.50 SD=1.14 40 —
Hispanic	M= -.03 SD=1.57 37 —	M= .22 SD=2.13 37 —	M=.86 SD=1.11 3 ↑ 33 — 1 ↓	M= .72 SD=2.16 1 ↑ 36 —
Asian	M=-.05 SD=2.75 25 —	M=1.40 SD=3.48 25 —	M=. 56 SD=.30 2 ↑ 23 —	M=2.43 SD=1.09 8 ↑ 17 —
Nonpoor	M=.52 SD=.95 1 ↑ 49 —	M= -.82 SD=1.26 49 — 1 ↓	M=.79 SD=.41 15 ↑ 35 —	M=.35 SD=.51 50 —
Poor	M=1.05 SD=.87 50 —	M=-1.48 SD=.80 50 —	M=.42 SD=.79 50 —	M=1.30 SD=.81 50 —
White-Black gap	M=.20 SD=.65 41 —	M=-.31 SD=.93 41 —	M=.13 SD=.33 40 —	M=-.66 SD=1.05 40 —
White-Hispanic gap	M=.62 SD=1.74 37 —	M=-1.22 SD=2.52 37 —	M=.20 SD=.71 37 —	M= -.86 SD=1.22 36 — 1 ↓
Poverty gap	M=-.48 SD=.25 50 —	M=.55 SD=.36 50 —	M=.50 SD=.72 50 —	M=-1.08 SD=1.36 50 —

Note. M is the mean of all states' growth parameter estimates, and SD is the standard deviation of the estimates across states. Statistical significance of each individual state's growth was determined at the .001 level to reduce familywise Type I error for simultaneous comparisons of multiple states. Numbers in front of arrows or dashes indicate the number of states for each pattern (↑ significantly up; ↓ significantly down; — no significant change).

Table C-7.
HLM Estimates of State Accountability Policy Effects on NAEP Grade 4 Reading and Math Trends (N=50 states)

		Reading		Math	
Group	Adjustment	Effect on Pre-NCLB Growth	Effect on Post-NCLB Change	Effect on Pre-NCLB Growth	Effect on Post-NCLB Change
All	Unadjusted	.11	-.13	.26**	-.45**
	Adjusted	.07	-.32	.20*	-.55
White	Unadjusted	.17**	-.43**	.21**	-.41**
	Adjusted	.14*	-.53*	.19**	-.51
Black	Unadjusted	-.11	.46	.06	-.16
	Adjusted	-.12	.43	.06	-.13
Hispanic	Unadjusted	.09	-.15	.08	-.17
	Adjusted	.10	-.01	.15	-.02
Asian	Unadjusted	-.03	-.39	.22	-.65
	Adjusted	.14	-.65	.46*	.05
Nonpoor	Unadjusted	-.04	-.06	.04	-.05
	Adjusted	-.01	-.09	.03	-.09
Poor	Unadjusted	.04	-.04	.34*	-.54*
	Adjusted	-.13	-.02	.15	-.41
White-Black gap	Unadjusted	.20	-.83*	-.02	.09
	Adjusted	.21*	-.01	-.03	.06
White-Hispanic gap	Unadjusted	-.03	-.18	-.11	.10
	Adjusted	-.02	-.19	-.11	.30
Poverty gap	Unadjusted	-.10	-.02	-.28**	.41*
	Adjusted	.11	-.10	-.06	0

Note. Numbers in "unadjusted" rows show estimated effects of state accountability policy without any statistical control for other covariates. Numbers in "adjusted" rows show estimated effects of state accountability policy with statistical control for initial status and pre-NCLB growth rate. Asterisks indicate statistical significance level of the estimate: * $p < .05$; ** $p < .01$

Table C-8.
HLM Estimates of State Accountability Policy Effects on NAEP Grade 8 Reading and Math Trends (N=50 states)

		Reading		Math	
Group	Adjustment	Effect on Pre-NCLB Growth	Effect on Post-NCLB Change	Effect on Pre-NCLB Growth	Effect on Post-NCLB Change
All	Unadjusted	-.12	-.17	.23**	-.14
	Adjusted	-.13	-.28**	.21**	-.15
White	Unadjusted	.05	-.39	.25**	-.27
	Adjusted	.04	-.31**	.20**	-.05
Black	Unadjusted	-.08	-.27	.07	.33
	Adjusted	.02	-.37	.08	.50
Hispanic	Unadjusted	-.44	.25	.56**	-1.23**
	Adjusted	.16	-.35	.35**	-.28
Asian	Unadjusted	-.94	1.27	-.44*	.73
	Adjusted	.29	.27	-.39	-.25
Nonpoor	Unadjusted	.01	-.30	.26*	-.20
	Adjusted	-.01	-.29**	.13	-.15
Poor	Unadjusted	-.24	.15	.59***	-.60*
	Adjusted	-.15	-.18	.18	-.17
White-Black gap	Unadjusted	.15	-.12	.10	-.68*
	Adjusted	.05	.10	.09	-.41*
White-Hispanic gap	Unadjusted	.34	-.47	-.35*	.96**
	Adjusted	-.02	-.02	-.28**	.45
Poverty gap	Unadjusted	.23	-.42	-.40**	.47
	Adjusted	.16	-.07	.04	-.29

Note. Numbers in "unadjusted" rows show estimated effects of state accountability policy without any statistical control for other covariates. Numbers in "adjusted" rows show estimated effects of state accountability policy with statistical control for initial status and pre-NCLB growth rate. Asterisks indicate statistical significance level of the estimate: * $p < .05$; ** $p < .01$; *** $p < .001$

Table C-9.
Correlations of State Accountability Variable with the NAEP vs. State Assessment Discrepancies in Proficiency Levels and Gaps by Grade and Subject

	Grade 4		Grade 8	
	Reading	Math	Reading	Math
All	.13	.31**	.17	.36**
White	-.02	.13	.05	.26*
Black	.15	.30*	.13	.32*
Hispanic	.02	.08	.14	.25
Asian	-.30*	-.33*	-.11	-.02
Nonpoor	-.07	-.12	-.00	.02
Poor	.20	.24	.29*	.35**
White-Black gap	-.27**	-.34**	-.14	-.36**
White-Hispanic gap	.02	-.03	-.07	-.10
Poverty gap	-.17	-.14	-.07	-.11

Note. The above correlation coefficients show the direction and strength of linear relationship between two variables: (1) the level of state activism in test-driven accountability and (2) the size of discrepancies between NAEP and state assessment in proficiency rate. Positive values mean the variables change in the same direction, whereas negative values mean they change in the opposite direction. Asterisks indicate statistical significance level of the correlation estimate: * $p < .05$; ** $p < .01$

Table C-10.
State Assessment vs. NAEP Discrepancies in 2003-05 Proficiency Gain by Grade and Subject (N = 25 states)

	State Assessment			NAEP		
	2003 % at/above Proficient	2005 % at/above Proficient	2003-05 % Gain	2003 % at/above Proficient	2005 % at/above Proficient	2003-05 % Gain
Grade 4 Reading	66.6	71.8	+5.2	30.3	30.6	+0.3
Grade 4 Math	61.4	67.4	+6	31.6	36.1	+4.5
Grade 8 Reading	63.1	66.6	+3.5	31.2	29.7	-1.5
Grade 8 Math	48.8	56.0	+7.2	28.8	27.9	-0.9